Received 20 June 1969 9.10, 9.1

Whither Speech Recognition?

Letter to Editor
J.Acoust.Soc.Am.

J.R. Pierce
Bell Telephone Laboratories, Inc., Murray Hill, New Jersey 07971

Research field of "mad inventors or untrustworthy engineers"

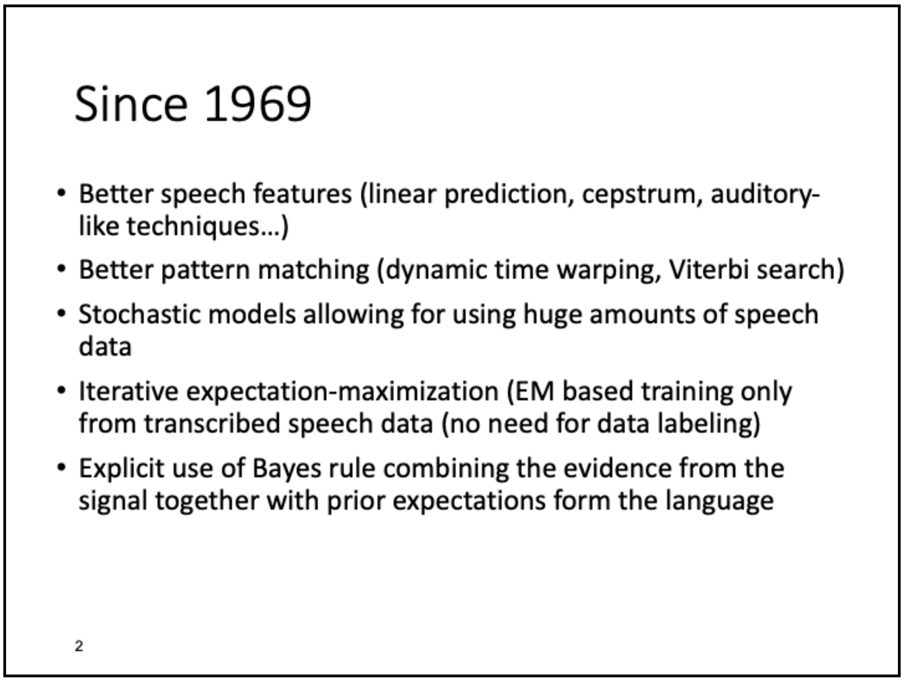
Funding artificial intelligence is real stupidity"

- supervised the Bell Labs team which built the first transistor
- President's Science Advisory Committee
- developed the concept of pulse code modulation
- designed and launched the first active communications satellite

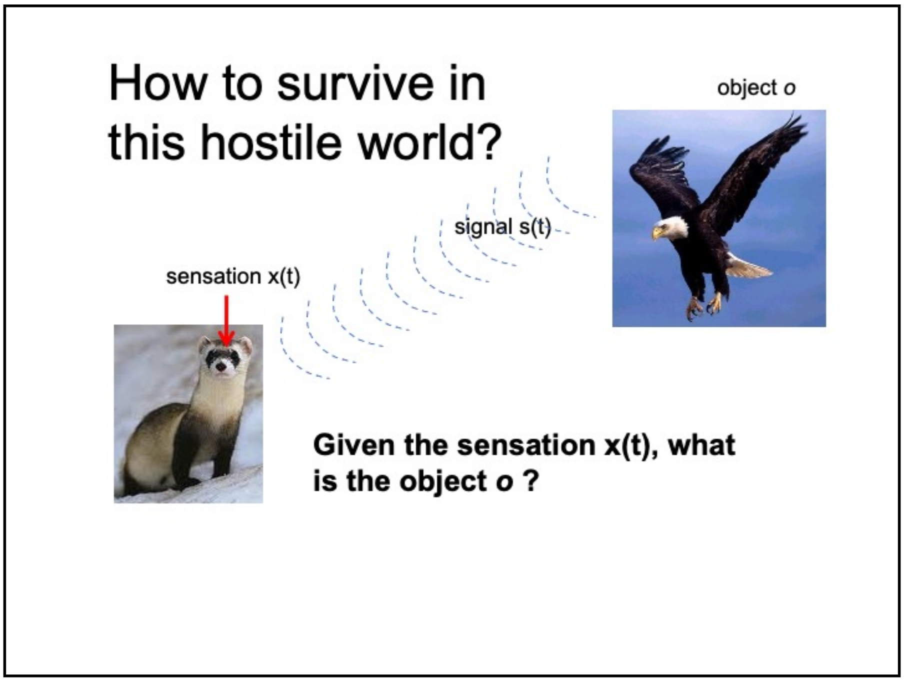
…. should people continue work towards speech recognition by machine ? Perhaps it is for people in the field to decide.

To implement ASR, we need to apply …. ***intelligence and knowledge of language comparable to those of a native speaker !***

A short letter to editor of the Acoustical Society of America from the very influential researcher at Bell Labs almost stopped speech recogntion research in USA. Read ithe letter by yourself, I believe that Dr. Pierce had some good advice, still valid even today.



The recognitions field fiortunately did not stop altogether and gradually it recovered to the point weher we are today. Several reasons for its recovery are listed here.



This is a situation, which is often faced in nature. Ferret's perceptual system receives a signal s(t). The signal may be multimodal (audio, visual, olfactory, tactile,..). The signal causes sensation x(t) in the perceptual system of the animal. The task, which our ferret need to perform is to find out what the sensation x(t) represents. In tis case, our ferret is concerned that the stimulus and the resulting sensation might have been produced by an eagle **o.**

Given the signal x(t), what is the stimulus s(t) ?

s(t) – e.g. changes in acoustic pressure representing the sound

x(t) – e.g. the activity in the auditory system

s(t) comes from some probability distribution P[s(t)]

- different stimuli have different probabilities of occurrences

x(t) occurs with the conditional probability P[x(t)|s(t)]

- x(t) depends on s(t) but the response is not unique (system noise?)

Need P[s(t)|x(t)] (probability that s(t) happened when the activity in the system is x(t)

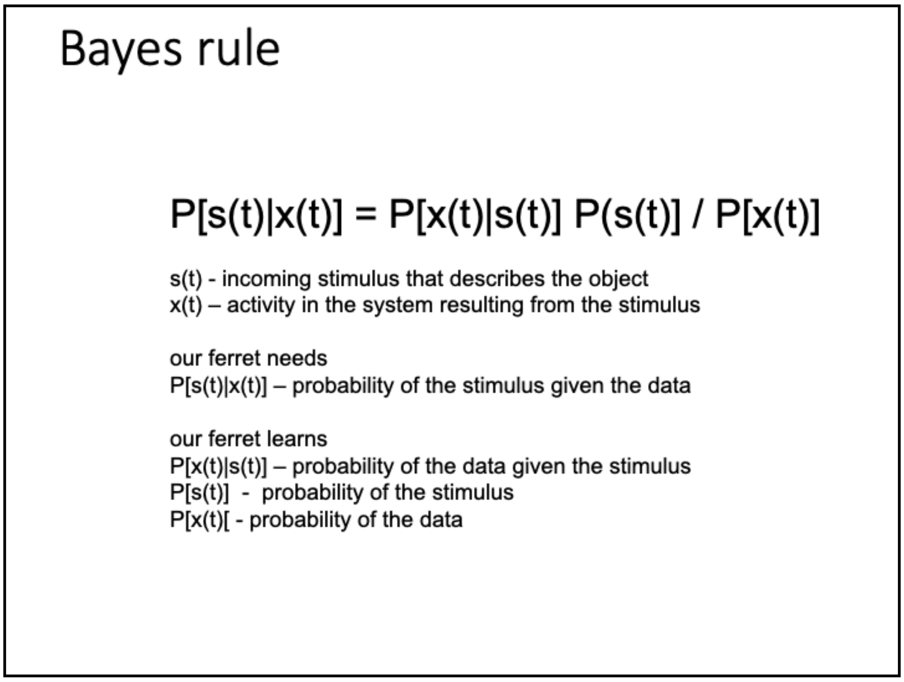
P[x(t),s(t)] – likelihood that both s(t) and x(t) happen at the same time

- have we experienced x(t) when s(t) happened ?

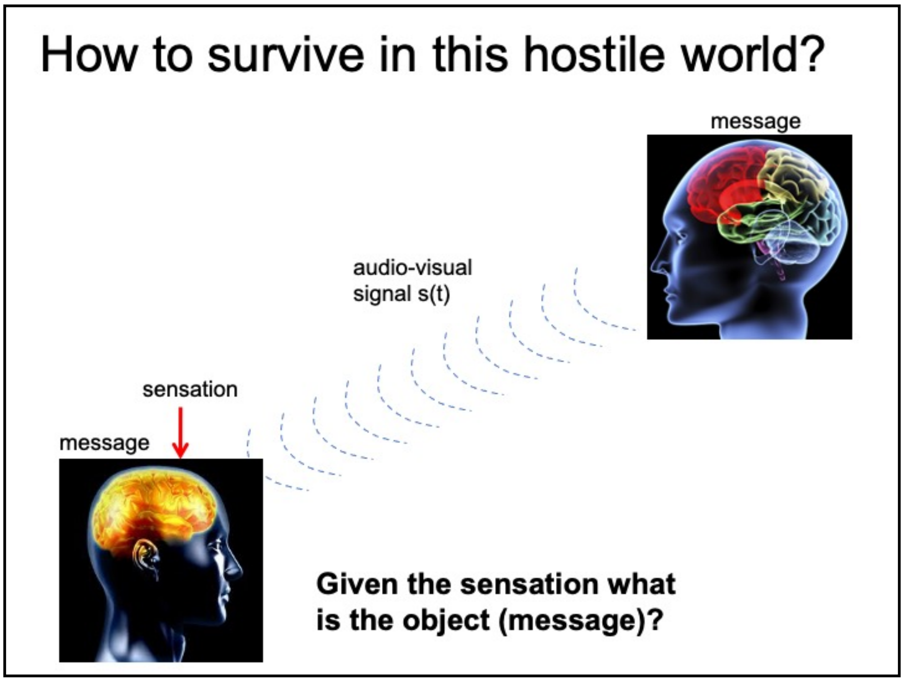
P[x(t),s(t)] = P[x(t)|s(t)] P[s(t)]   or   P[x(t),s(t)] = P[s(t)|x(t)] P[x(t)]

P[s(t)|x(t)] = P[x(t)|s(t)]P(s(t)]/P[x(t)]

(Bayes Rule)



This is just to summarize what we have talked about before. What our ferret needs is to estimate what is the probability P[s(t)|x(t)] of the eagle when hearing the eagle's cry x(t). The ferret knows from its prior experience what if the probability of eagle cry given the eagle is in the sky P[x(t)|s(t)], probability of heaving eagle in the given context P[s(t)], and also what is the probability of hearing the eagle cry x(t).



Humans developed different acosutic means to survive, the communication by speech. Using speech, besides an obvious means for issuing warnings (many animale can do the same using sounds), an amazing amounts of information can be exchanged. In particular, we belive that speech is unique among communication means since it is able to be used to communicate abstract concepts (truth, justice,…) or to refer to past or future. It is tempting to say that speech is what differetiates humans from other species.

Stochastic machine recognition of speech

$$\boldsymbol{P(M,x) = P(M|x)P(x) = P(x|M)P(M)}$$

Joint probability that message *M* and data *x* occur together is given by the probability of the message *M* given the data *x* multiplied by the probability of the data *x*, or by probability of the data ***x*** given the message *M* multiplied by the probability of the message *M.*

Bayes rule

$$P(M|x) = P(x|M)P(M)/\ P(x)$$

To find the maximum of the probability, the probability of the data is not need

$$\boldsymbol{M = \ \textbf{argmax}\, P(x|M_i)P(M_i)}$$

How to get good model M ?

How to find the optimal *M*?

What is the data *x* ?

Formarly, we express the join probability of the particular sequence of models (the message M) and the observed data. Joint probability that message *M* and data *x* occur together is given by the likelihood of the message *M* given the data *x* multiplied by the probability of the data *x,* or by probalikelihood of the data ***x*** given the message *M* multiplied by the probability of the message *M.* This yileds so called Bayes rule, which expresses the liklelihood of a given message *M* given the data *x*. Since the probabilty of data is just a scaling factor (does not depend of the message M) it can be ignores. So to find the message M which most lilely generated the data x, we need to search all possible messages which could be generated by the model and keep the ine which yields the highest likelihood.

Easier said than done, since the model M may not be correct, the acoustic model lilelihoods (learned from the acoustic training data) may be inaccurate,, the prior probabilities of particular messages (the language model traind on text data) may be also incorrect,

the number of possible messages is huge and searchin through all of them is not easy, and the features x derived from speech signal may not carry the required information and may carry information about a number of irrelevant information soutrces.

How to get good models ?

The model architecture and **DATA !**

$$\boldsymbol{M} = \ \mathbf{argmax}\, \boldsymbol{P}(\boldsymbol{x}|\boldsymbol{M_i})\boldsymbol{P}(\boldsymbol{M_i})$$

There is no data like more data
attributed to Bob Mercer

More data is always better than more thinking
attributed to Eric Brill

No knowledge is better than wrong knowledge

To get good model of speech, we need good model architecture and good data for the model training. The architecture largerly stabilized as a sequence of speech sounds. What represents the is dependent on amount of training data which is available for the model training. More detailed models of sounds require more data. Similarly, more complex language models may require more data. So to get better models needs more data,

How to find the best *M(w)*?

Search through all possible models $M(w_i)$ to find which most likely produced x

$$M(w) = \underset{i}{\mathbf{argmax}}\, P(x|M(w_i))P(M(w_i))^{\gamma}$$

ϒ – "fudge factor" (makes all statisticians uneasy)

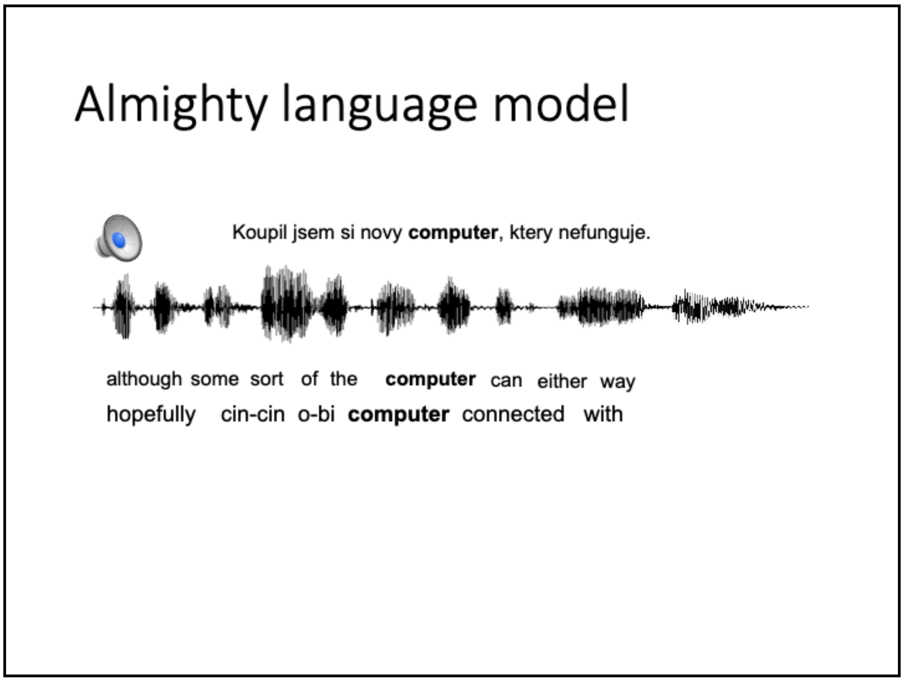
The search is a form of dynamic time warping where templates are represented by stochastic models. Where distances are replaced by likelihoods of $P(x|M)$ combined with priors $P(M)$

$P(x|M)$ from acoustic data (labeled by speech sounds

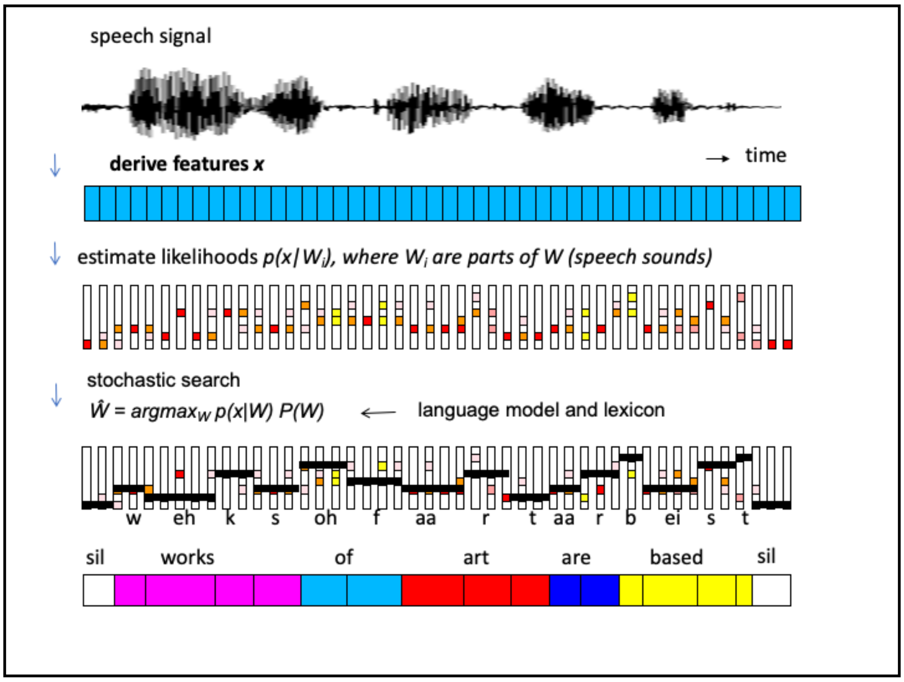
- trivial when sound labeled training data
- not labeled but only transcribed
  - know the sound sequences, find boundaries through iterative expectation-maximization techniques (remember the line of boys and girls)

P(M) – sequences of words with probabilities derived most often from texts
  - Problems for low or zero probability words!

The search for the best message involves two stochastic models, the acoustic model *P(x|M) which is trained using labeled speech acoustic data, and the language model P(M), typically trained usimg the text data. Realtive contributions of these two models is controlled using the "fudge factor"* ϒ, which is chosen experimentally for the best performance on some development data. Notice that messages for which the *P(M)* is zero cannot ever be chosen. Thus, the words which are the most likely in the language are also the most likely be selected. Remembering that from the information point of view, the less likely items carry more information, we see some inconsitence.

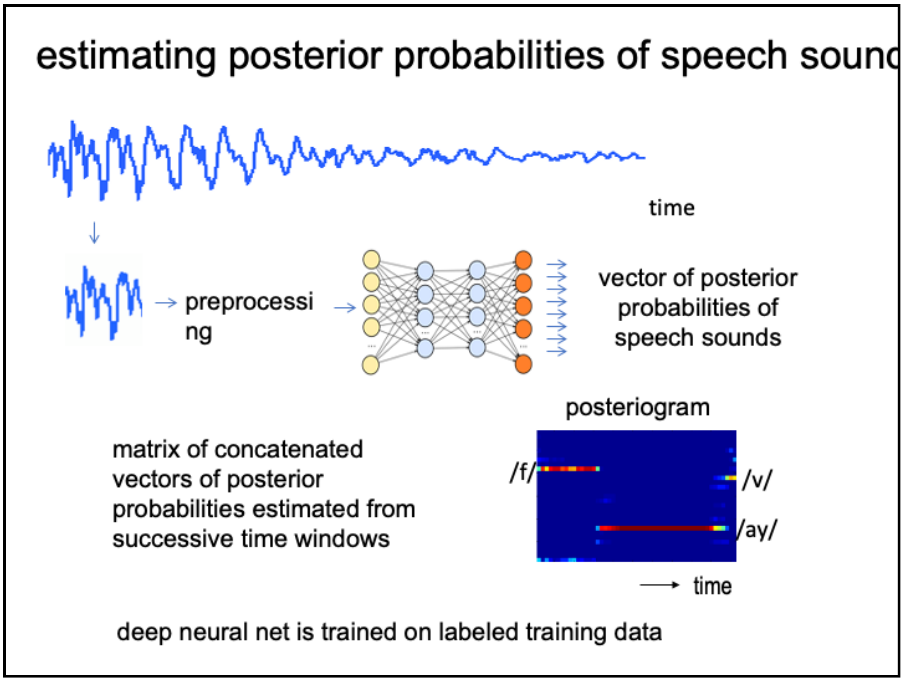


Language model can severely limit what can be recognized. THis is obvious when you try to recognize a sentence in a language which your recognizer does not expect. Two outputs from two different sate-of-the-art systems available today in response to Czech language are shown here.

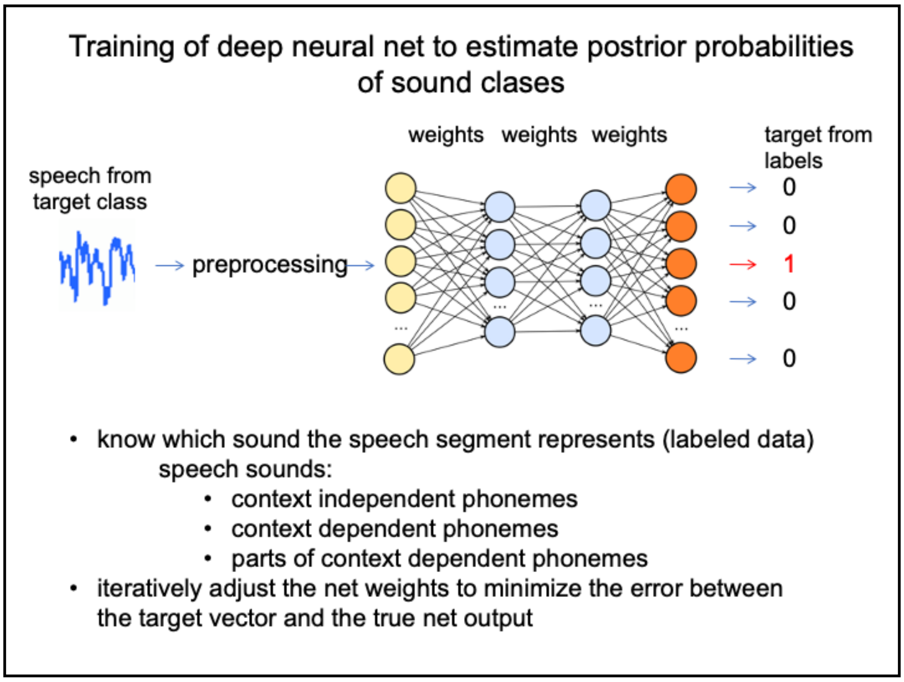


This scheme shows schemitacally the HMM recogntion process. Short speech segments (what is "short" may be discussed later) are described by the segment features ***x***. Likelihoods of speech sounds given the features are derived using trained acoustic models. Search for the best path though the sequence of likeligood vectors with further help of the language model, which contains preferences for a particular word combinations, yields the sequence of speech sounds (phonemes). Having this sequence, pronunciation rules allow for transcribing the unknown utterance/

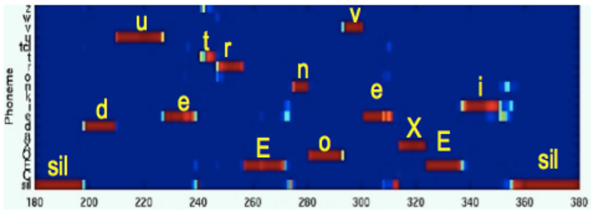
Deep Neural Nets in Speech



DNNs can be used for deriving posetrior probabilities of speech sounds. Input is a set of some features describing the informatipn in the short speech segment (think short-time spectrum of the windowed speech segment), on the output of the trained net we obtain estimate of proterior probabilties of all speech sounds in which the DNN was trained.



For their training we need again labeled speech data. The DNN is trained to deliver hghest output for the sound which undelines the given segment of the speech signal. This is done using the "one-hot (that is one output is close to 1, others are close to 0) vestors as the target of the DNN during the traing. The weighst of the DNN are iteratively adjusted during the training to achieve this as well as possible.



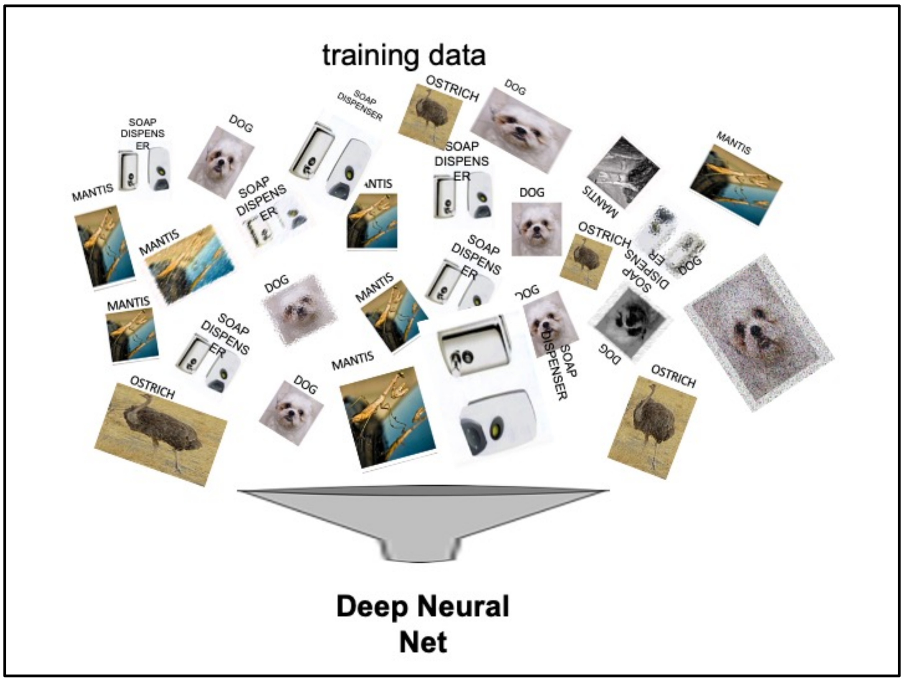
Neural network classifiers

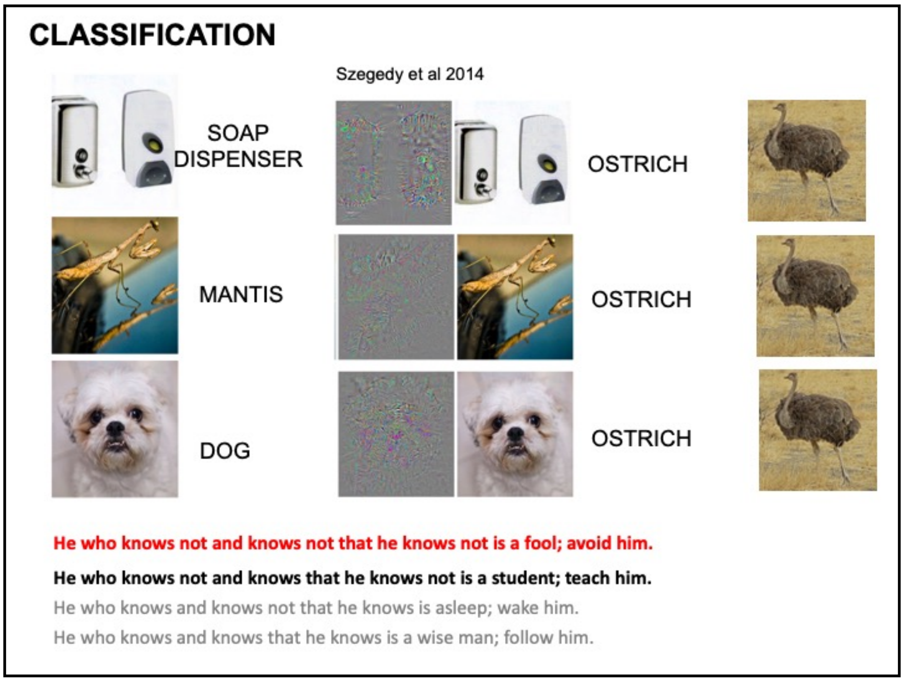
good: can be very accurate on arbitrary distributed and correlated data

bad: need a lot of training data

ugly: can be very wrong without much warning

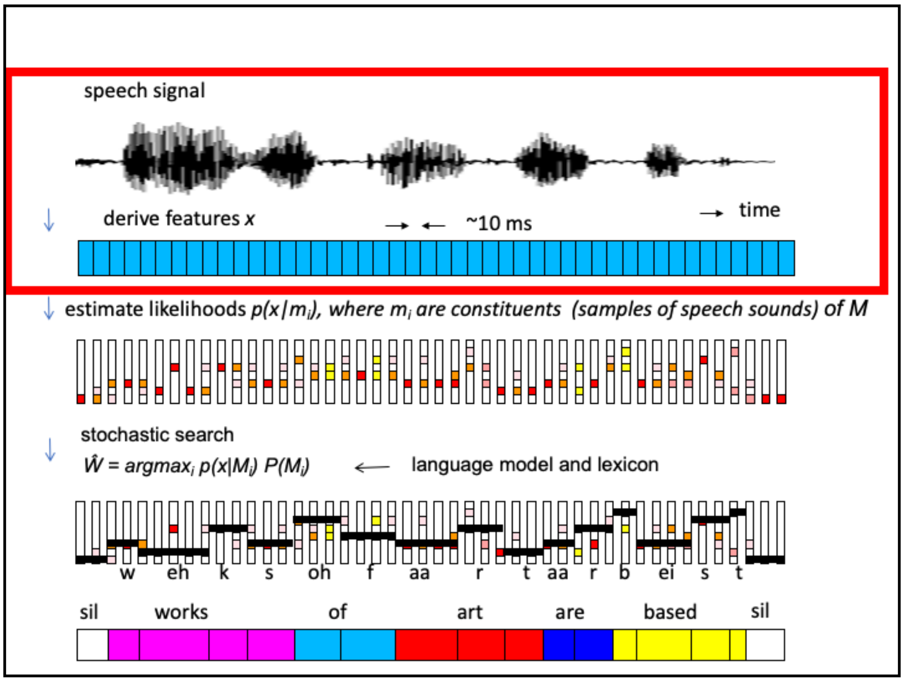
Well trained net on good clean data can estimate posterior probabilities of speech sounds amaizingly well. As inputs DNNs do not require uncorrelated and Normally distributed data, the main requirement typically is that the data are “whitened”, i.e. mean and variance normalized. For a good performance, DNN need significant amoounts of training data. This also implies that during the inference (diring recogntion) they can be very sensitive to any deviations from the statistics of the training data.


training data
SOAP DISPENSER
OSTRICH
DOG
MANTIS
Deep Neural Net



How bad can DNNs be in classification? This is nicely illustrtated in examples of classification of still images. When the signal is clean, the classifier works perfectly, **soap dispenser, praying mantis** and **dog** are classifier as such.

When a small amount of noise, not perceived by human observers at all, is added to the images, all images are classifed as **ostrich.** No surprise, this bothers at least some researchers.



We have spent time discussing hot to derive a good set of features **x** for ASR. As long as the likelihood estimates were derived by the quadratic diagonal-covariance Gaussian mixtures classifier, the features ***x*** needed to be approximatly NOrmally distributed and uncorrelated. That limited the techniques which were used for their estimation. There is no such a contstraint for the deep neural net (DNN) classifiers.

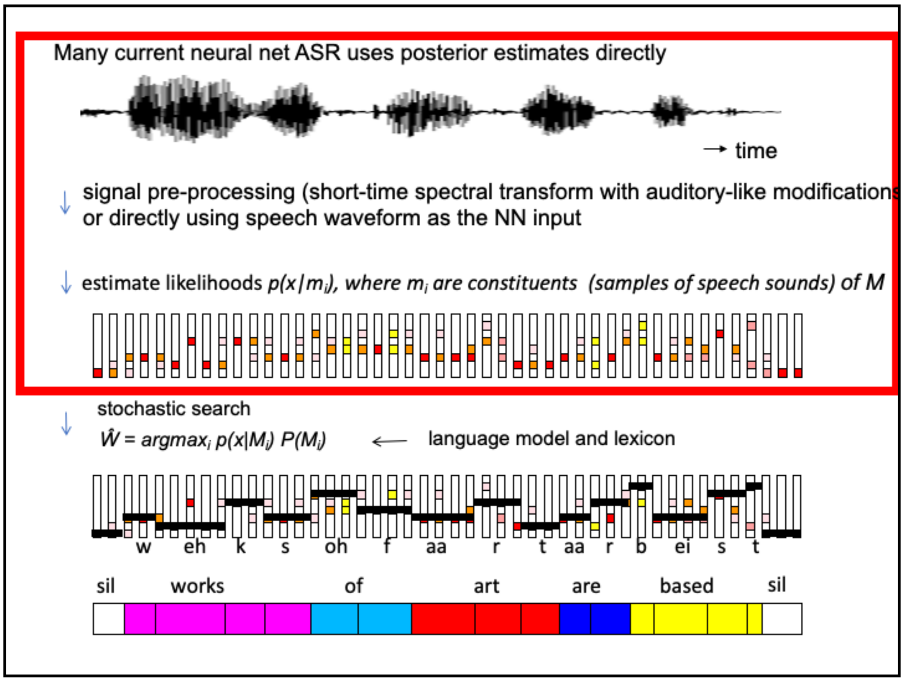
DNN-HMM hybrid

- Convert posterior probabilities to likelihoods (divide by training priors) to be used in Viterbi search for the best word sequence

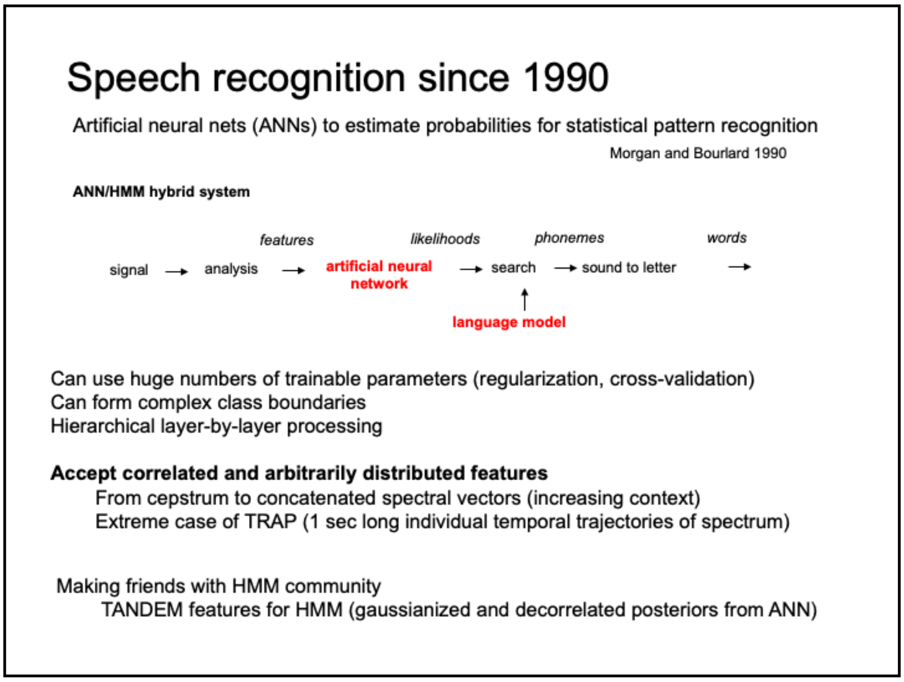
Bourlard and Morgan, NIPS 1990

$$p(c_i\,|\mathrm{x}) = \frac{P(c_i\,|\mathrm{x})}{P(c_i)}$$

One way of using posterior probability estimates in ASR is to forget about the classical Gaussian classifiers used for the sound class likelihood estiamtes in HMM paradigm since the seveties of the last century and use the postrior probabilities from the DNN directly. To convert the postrior probablities to likelihoods for the HMM search, we can divide each postrior estimate by the class prior (which is known from the training data). Today is sounds easy and typically works well but in the early days of DNNs, estimating probabilities of large sets of sound classes which were used in ASR, was not easy. Additionally, GMM-HMM recognizers were well developed and it was not easy to abandon them.



Here is the simplified ASR system, which directly uses estamates of likeligoods of speech sound from the DNN.



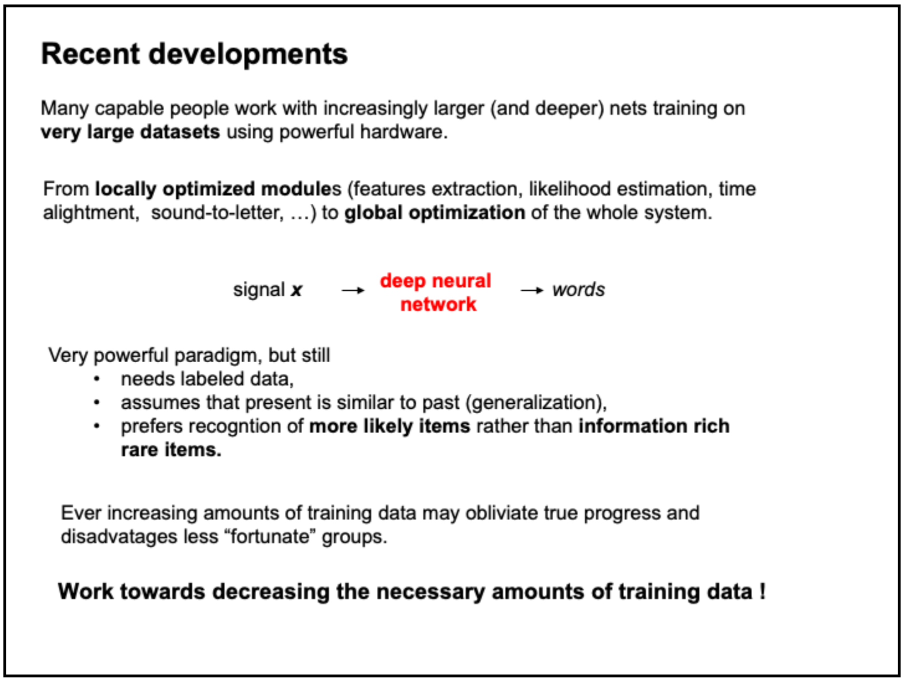
Late eighties and early nineties brought back the Rosenblatt's perceptron, this time with nonlinearities on each processing layer and with straigtforward adaptation of net weights.

Initially, beasuse of the lack of data and computing resources, it was difficult to use large ANN models and the nets were competetive only in simple archtectures such as single-state phoneme model based ASR.

However, the ANN advanteges such as the ability of forming a complex class boundaries, hiearchical layer-by-layer processing, and finaly the ability to deal with wide class of input features, were always there.

It encourage us to inttroduce larger and larger temporal contexts, to abandon the cepstrum based envelope for the individual spectral energy trajectories, sometimes as long as 1 second,

ANN people also contributed to the competetive GMM/HMM sustems by providing them TANDEM features derived by ANN, which were fordced by ETSI (European Telecommunications Standards Institute) requirement of using HMM/GMM based classifier in the competition.



I would like to stop this brief and necessarily incomplete history description at the point where most papers in this conference will probablyt be,

that is at the current explosive development in use of ANN in their current deep-learing form.

Many very competents and capable researchers are introducing complex

huge net archirtectures trained on unbelievable amounts of training data using more and more powerful hardware.

The most important advance is that the module-by-module optimization is gradually replaced by the global optimozation.

The end-to-end approaches take the input data, sometimes even in the original raw signal form and directly deliver a sequence of letters describing the message.

Still, some issues from the early days of ASR remain: the need for some labeled data is here, the assumption that the operation domain is similar to the domain on which the system was trained,

and the preference of frequenct and therefore expected words ovwer the information rich infrequent words is still here.

The need for training data is still exponentially increasing and the system performance is increasing with it.

We may not know anymore if the improvements come from some fundamental improvement of the underlying principles or from more data.

~ 300 days of labeled data

~ 100 years of unlabeled data

Parthasarati et al 2019

137 billion free parameters

100 billion words in training

Shazer et al 2017

What should be the signal x ?

No knowledge is better than wrong knowledge
but **more we know, less we need to learn.**

We speak in order to be heard and need to be
heard in order to be understood.
Roman Jakobson

Features *x* (early decision making)
Alleviated relevant info is lost forever, irrelevant info that is left in may create problems during the inference.

Emulate some relevant properties of human hearing in feature extraction ?

More we know. less we need to learn.

And what we know for sure is that speech is

perceived through hearing. So we initially decided to emulate some basic properties of human hearing in the feature extraction module of ASR systems

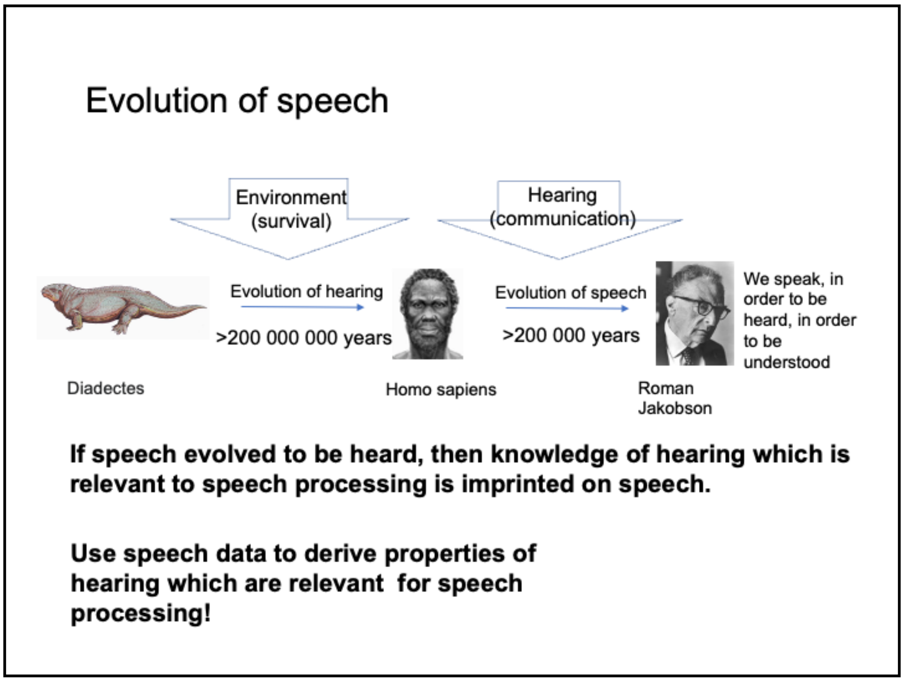
The feature module was one of obvious points in the system to start with.

Features need to represent linguistic information about message in speech and not irrelevant information (speaker specific, acoustic environment,…)

whatever is alleviated during feature

computation, is lost forever

whatever is left in, makes the recogntion (training) more difficult



Large amounts of labeled speech data are becoming available. It would be a shame to use it only for bling training of ASR algorithms.

Should Airplanes Flap Wings ?

"Airplanes do not flap wings but have wings nevertheless,…..

Of course, we should try to incorporate the knowledge that we have **of hearing, speech production, etc.**, into our systems,….but we need to estimate the parameter values from the data. There is no other way

F. Jelinek, Five speculations (and a divertimento) on the themes of H. Bourlard, H. Hermansky, and N. Morgan, Speech Communication 18, 1996, pp. 242–245

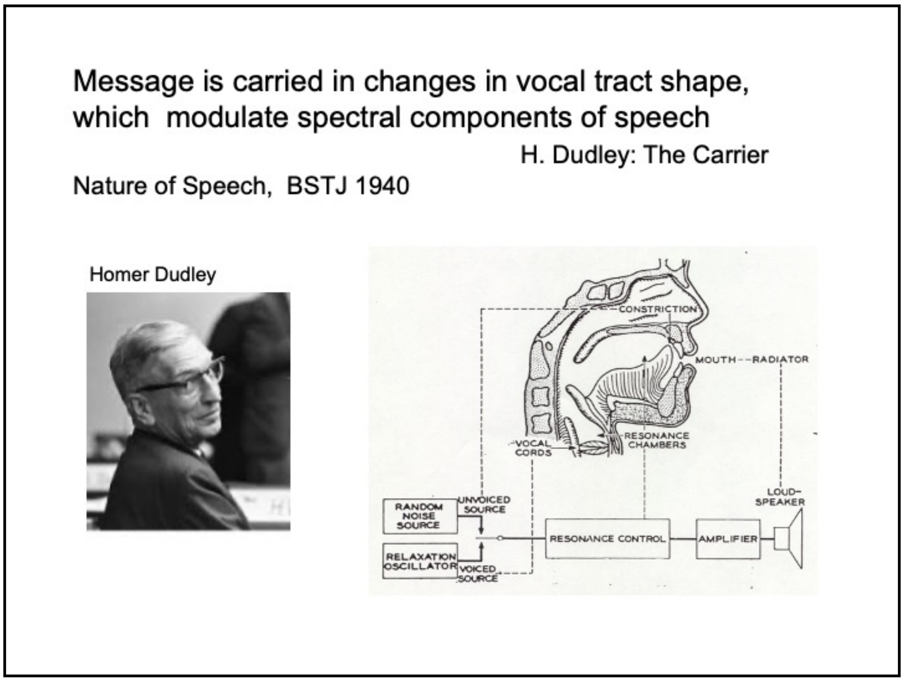
We have seen that emulating some basic properties of hearing such as nonequal spectral resolution of realtive insensitivity to unchanging spectral patterns can be usefol in ASR.
However, there is always a question which properties are relevant for speech and which serve other hearing purposes (pr no purposes at all, such as appendix).
Fred Jlinek correctly suggests that properties that are emulated should be supported by speech data.
Indeed, large amounts of labeled speech data are becoming available. It would be a shame to use it only for bling training of ASR algorithms.

Spectral Envelope

Speech Production

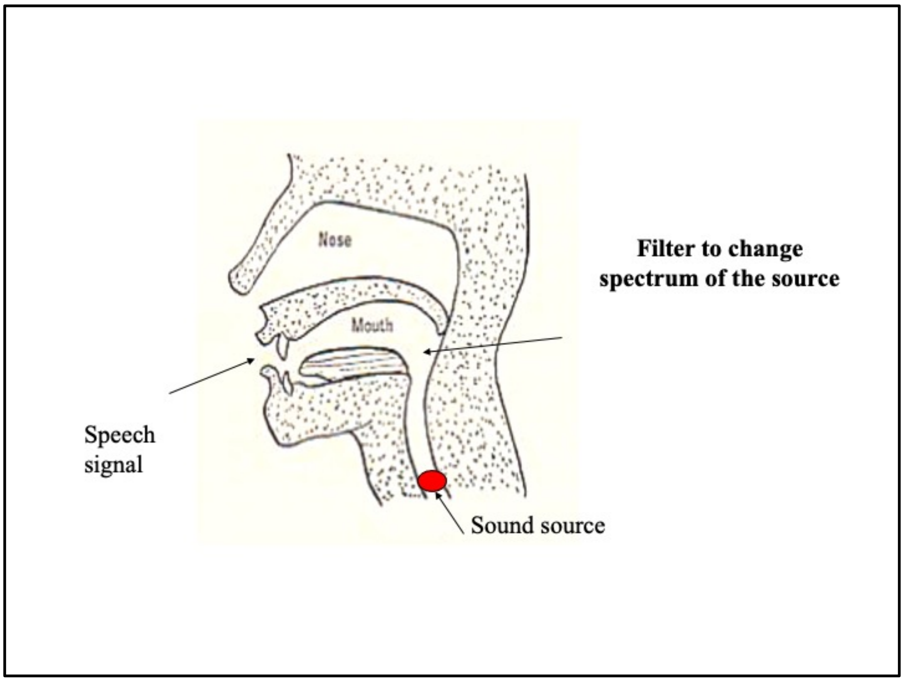
27

The messages are carried in movements of vocal tract of a speaker. Ther movements would not be heard but humankind invented means for making them to be heard by exciting the vocal tract by sound sources with rich audible spectrum and realtive flat spectral envelope - the vibration of vocal cords, air frictions in narrow passes in the tract (including the non-vibrating vocal chords in production of whicpered speech) or sound pulses in release of aor in plosive soiunds. These roch source spectar are modulated by changong transfer function of the vocal tract. This the message information is carried in the canges of the audible audible spectra of speech.

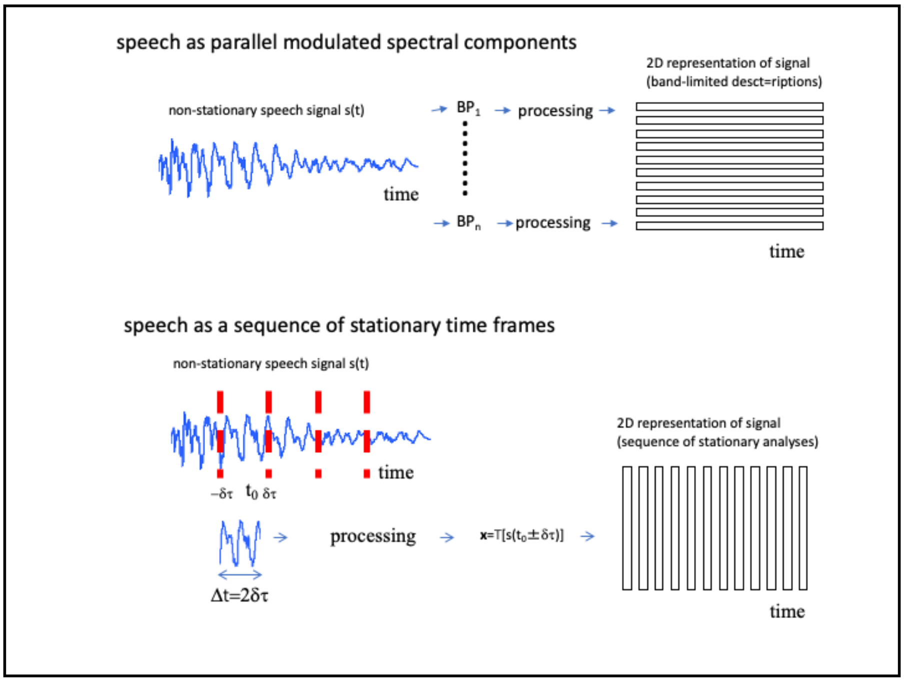


One important concept was the concept of carrier nature of speech. It says that the bulk of the information about the message in speech is in **changes of vocal tract shape,** i.e. in slow movements of

vocal organd such as tongue or lips. These movement are made audible by exciting the vocal tract by the source which is ring in overtones (vibration of vocal cords, air friction in tract costrictions, sudden release of air after the constriction release, ,,,). Noticethere was no mentioning of vocal tractr resonaces (formants) in Dudley's concept.

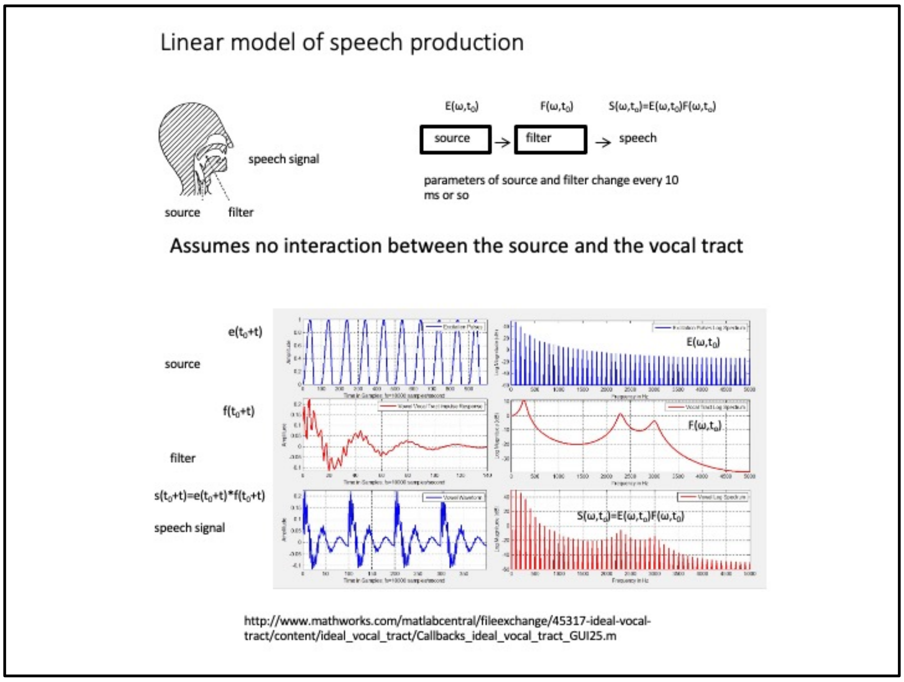


Two basic elements of the vocal tract are the **tract cavities** which filter the spectrum of the **sound source**. So far, all was fine. Spectrally rich source signal gets modulated by changing filtering properties of the tract shape. However, mathematical analysis is process in not easy. Some simplifi=ying assumptions may be needed.



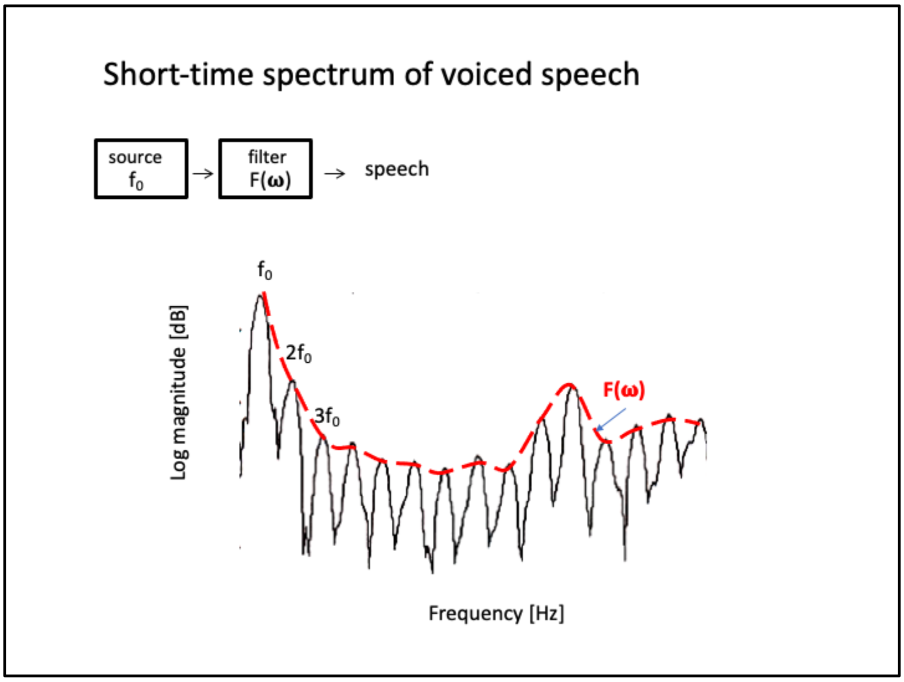
In the original concept (Homer Dudley 1940) the speech signal was seen as a number of parallel band-limted signals modulated by different parts of the changing frequency response of the moving vocal tract.

As digital signal processing techniques started to dominate speech processing, the concept gradually shifted to a time-frame processing where time segment of speech are are short enough so that the signal can be considered to be stationary. Parameters ere extracted from these segments independently of the other neghbousing segments. The 2-D time-frequency representaion of the signal is a sampled representation. The frame rate of the 2D representation implies the highest frequency of change preserved in this representation.



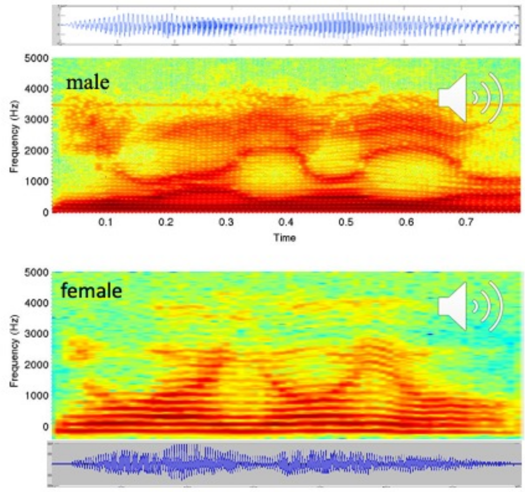
In early sixties of the last century in works of Gunnar Fant in KTH Stockholm and Kens Stevens at MIT studied extensively the naturu of both the source and the acoustic filter under the assumption that there is no interaction between them (hence the linear model). Some interaction of course exists - the filter gets dumped when the glotis in the source opens and the lung cavity gets into the action but this is mostly neglected. Further, the whole system is assumed to be stationary for short sedments of time (10-20 msec) – another assumption which is not entirely correct. However, these assumptions allow for cleaner mathematical formulation of the speech production.

DIgital techniques for short-time Fourier analysis sealed this mathematically clean but rather approximate concept, which mostly remains in speech processing until these days. Speech signal is first chpped into short speech segments, each segment is analysed independently of other segments, and these discretized analysis results are describing the dynamic speech signal.



Spectrum of voiced speech consists of spectral envelope $F(\boldsymbol{\omega})$ which characterize the acoustic filter and the fine straucture which characterizes the speech source. The filter has spectral peaks where the vocal tract resonates(formants). The fine structure has peaks smaced in integral multiples of the discributed in integral multiles of the fundamental frequency f_0 which is the frequency of the vibrarion of vocal chords.

Spectrogram from short-time fourier transform

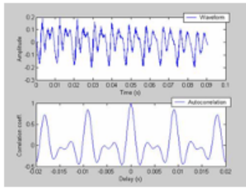


Spectrum from Fourier transform has equal spectral resolution at all frequencies source → filter → speech

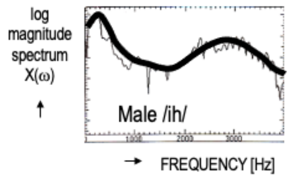
Source: Periodicity of the signal within the window

e.g., peak of signal autocorrelation in the window

$$r(\tau) = 1/2\delta t \sum_{k=j-\delta t}^{j+\delta t} s_k \, s_{k-\tau}$$

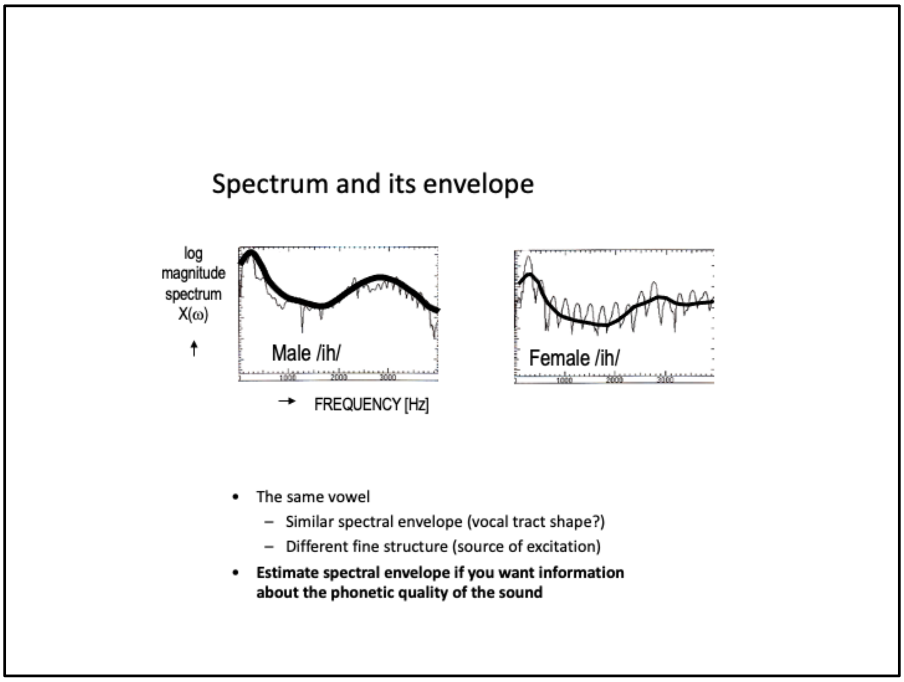


Filter: envelope of the short-time spectrum

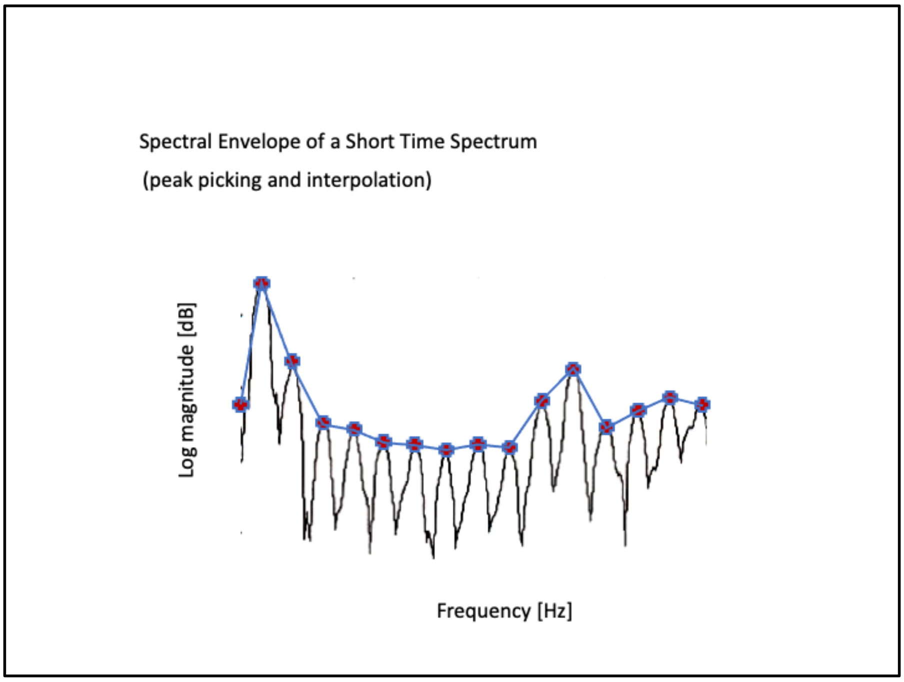


Message is in changes in **shape of a vocal tract** (change of the tract shapes the spectrum of the carrier)

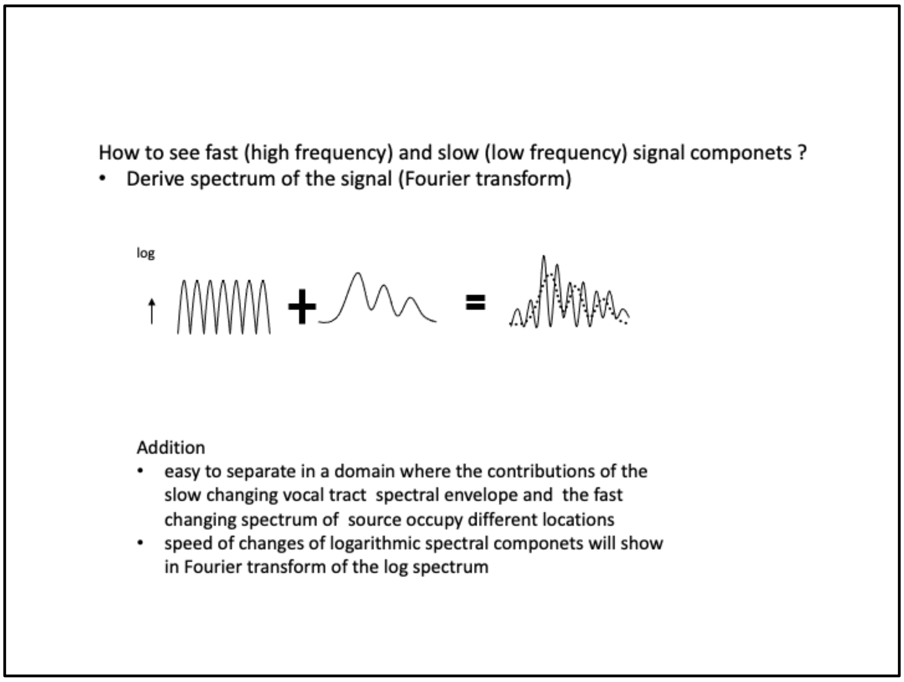
Simplest emulation of the voiced source can be characterized by a train of appropriately shaped pulses with the perion of the signal wavefors (can be estimatem by different means, the most starigtforward one is, e.g. , the signal autocorrelation. The filter is characterized by the spectral envelope of the short-time Fourier spectrum, the source is represented by the spectral fine tructure.



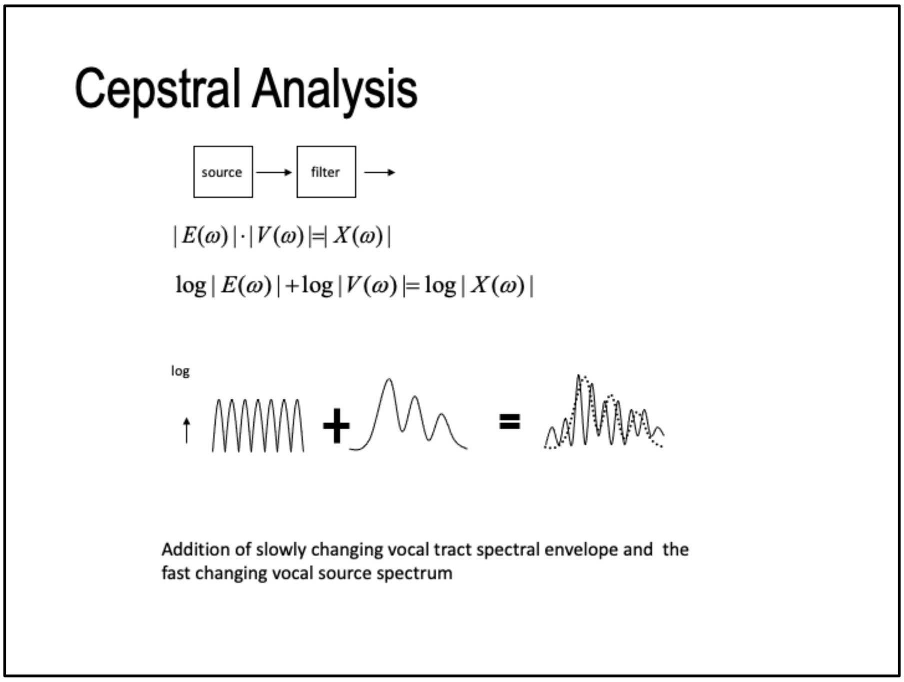
Fine structure of speech spetrum differs considerably among speakers of different genders, female speech has typically higher fundamental frequency ro that the fine structure components are more widely spread, male speech with its lower fundamental frequency has harmonic peaks closer to each other. However, spectral envelopes of speech segments with the same phoentic values are more similar among speakers.



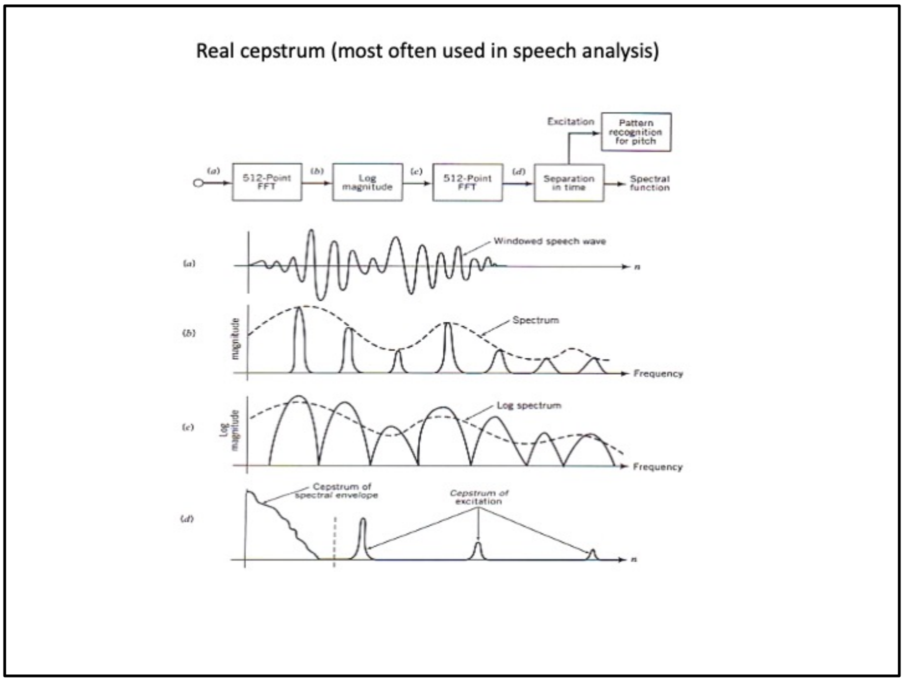
One straightforward way of estimating spectral envelope is to find tips of harmonic peaks (their are spaced in integral multiples of the fundamental frequency) and to interpolate between the found peaks.



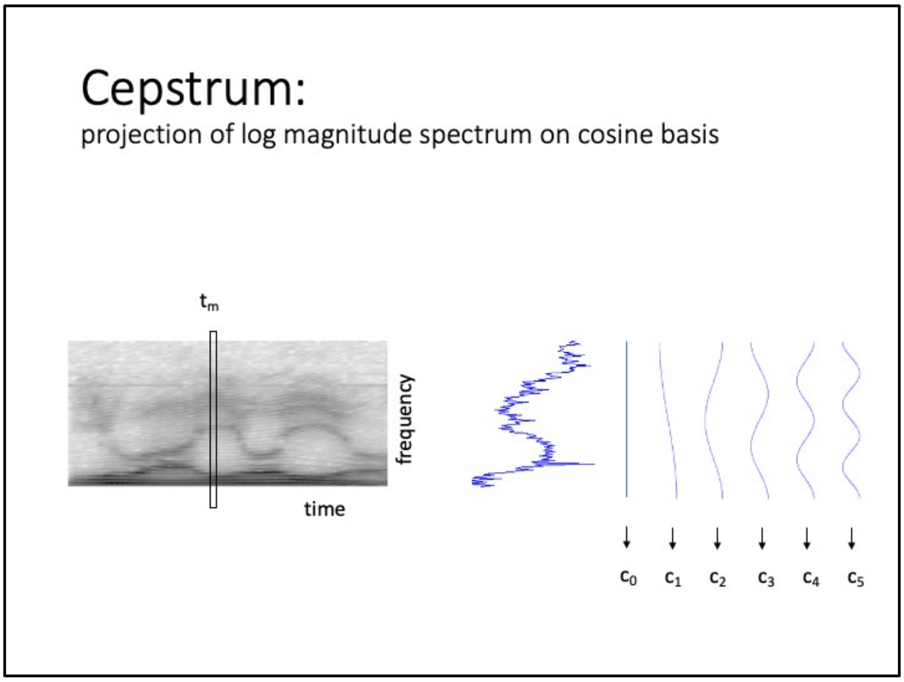
Logarthmic spectrum of speech is seen as a superposition of the flat spectrum of the source, which changes fast in frequency, and the slower changing spectrum of the spectral envelope. SIncet hese additive components of the logarithimc speech spectrum are changing with different rates in frequency, are more clearly separable whne taking fouirer transform of the logarithmic spectrum



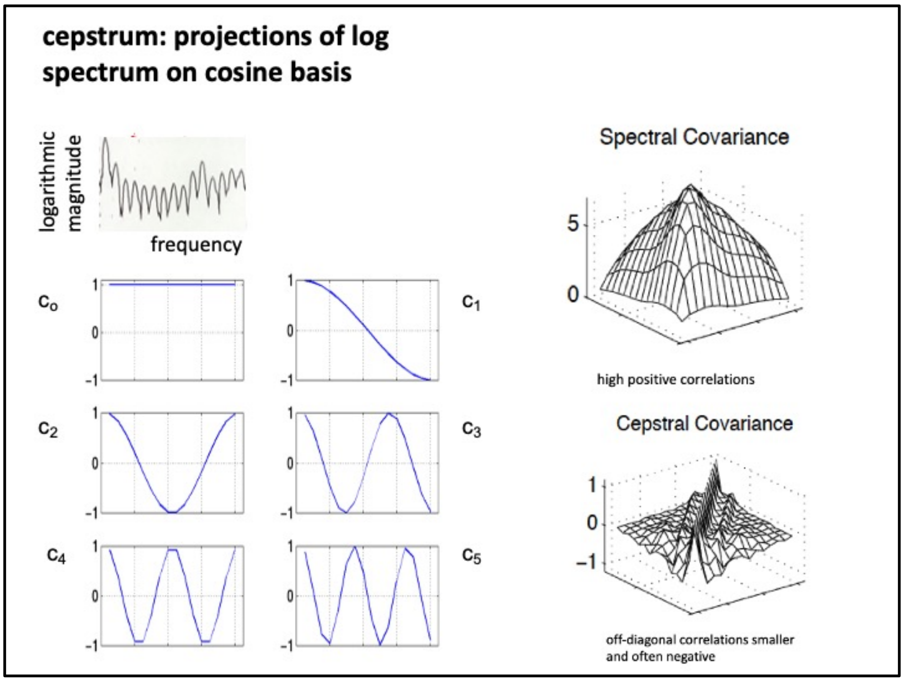
Mathematically, the log spectrum of speech is given as a sum of log spectra of the source and the spectral envelope



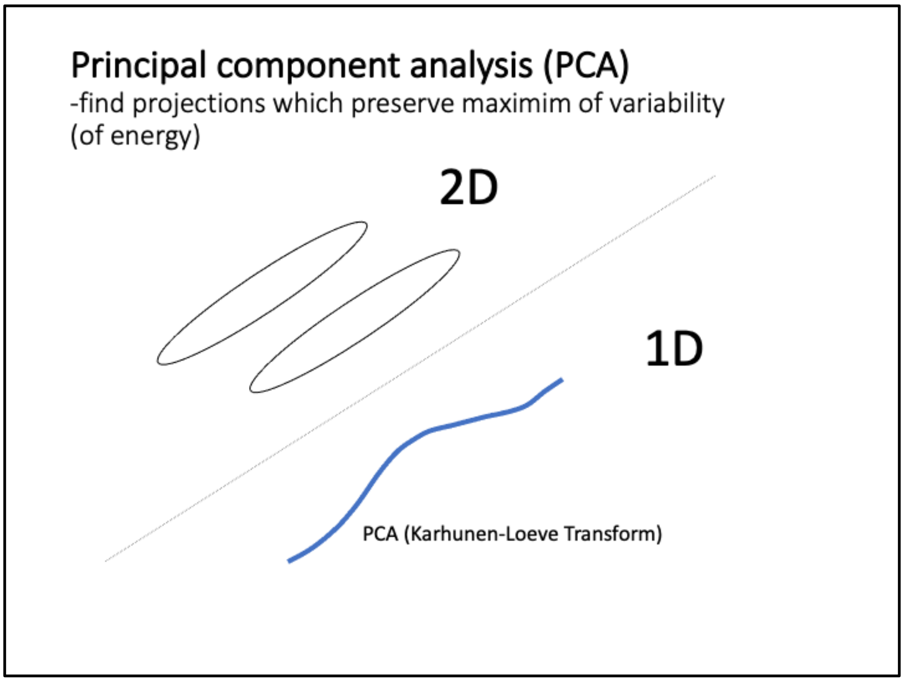
After taking fourier transform of the logarithmic speech spectrum, we obtain so called ***cepstrum*** where the contrution of the spectral envelope is at the low cepstral time coefficients and the contribution of the source are at higher cepstral coefficients.



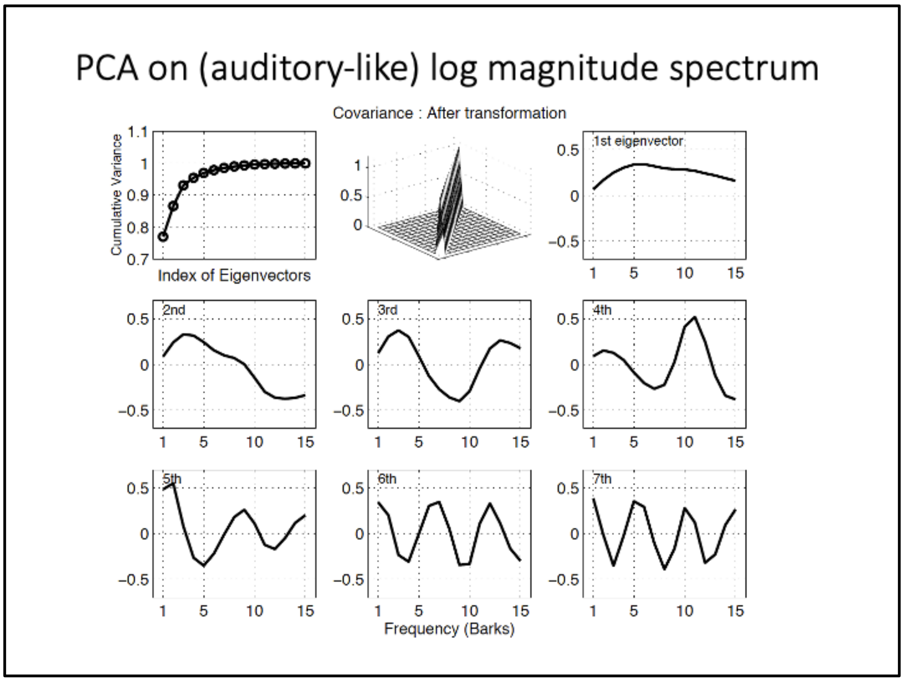
Cepscrum can be seen as projection of the short-time spectral on the bases formed by integral multiples of half period of cosine functions with decreasing period (increasing frequency)



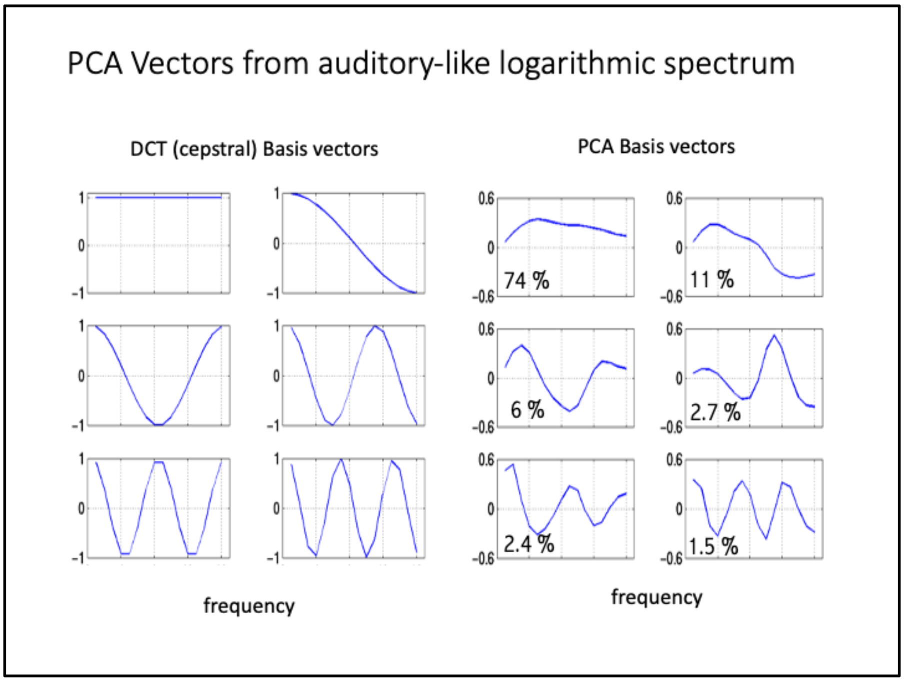
Another reason why the cpestrum can be useful (especially in pattern classifications) is that the cepstral projections approximately decorrelate the speech spectra. This allows for simpler (diagonal covariance) classifiers.



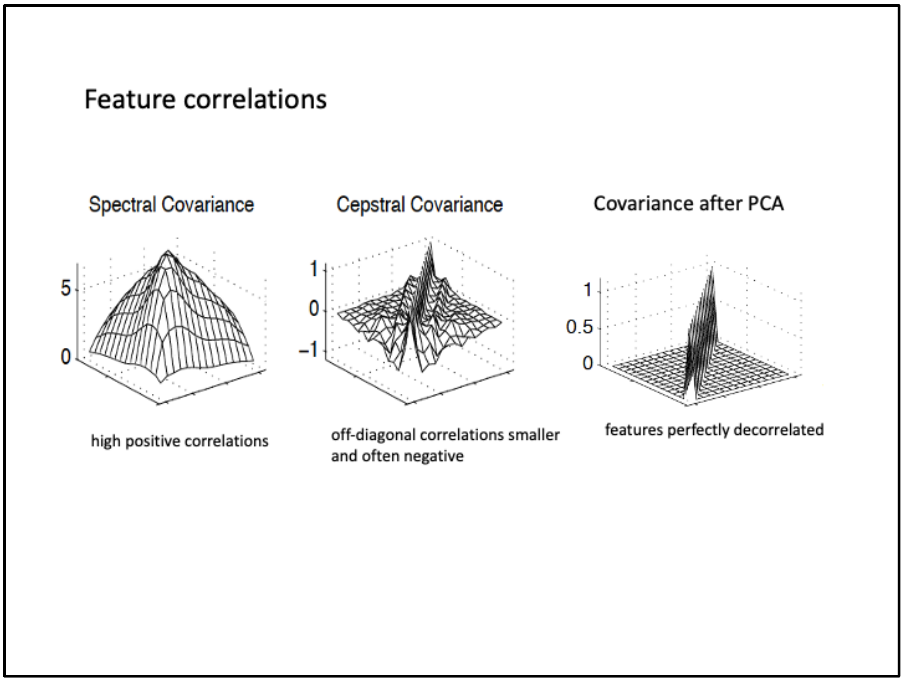
The perfect way of decorrelating vector space is to project the space in its principal componets. The can be easily derives by the principal component analysis (PCA) of the space, involving the computation of the covariance matrix of the space. PCA projects the space in directions of its maximum variability. If you believe that the information you areseeking is in its variability, PCA is a good way of reducing free parameters in your representation.



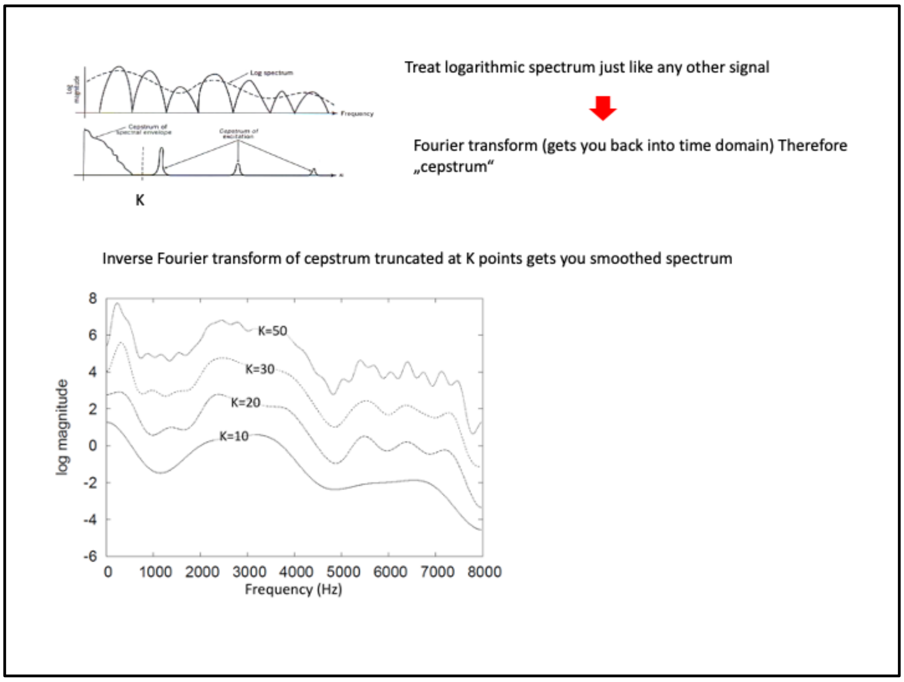
Computing PCA coefficients on (auditory-like to be discussed shortly) spectral space indeed yields projections which are similar to cosine projections in the cepstrum.



Comparing the cepstral and PCA projections, we see the similarites



And comparing the covariance matrices of the cepstrally and PCA projected log spectral spaces also shows that cepstrum is a reasonable “poor man” approximation of the PCA projections (without the burden of deriving the PCA bases).



Computing cepstarum of the logarithmic spectrum gets us to the time domain. By projecting the cepstraum back to the spectral domain using the inverse fourier transform gets us back to the logartihmic spectral domain. Depending on how many cepstral coefficients we keep for the inverse fourier transform, we obrail different amouts of smoothing of the original logarithmic spectru,=m.