\

Some history of acquiring knowledge about speech communication.

1

History

"Those who don't know history are doomed to repeat it."
– Edmund Burke

And sometimes it may be desirable ☺


DEVS
JEHOVA
Angeli
Stellæ
JEHOVA

How does hearing get the information about speech sounds ?

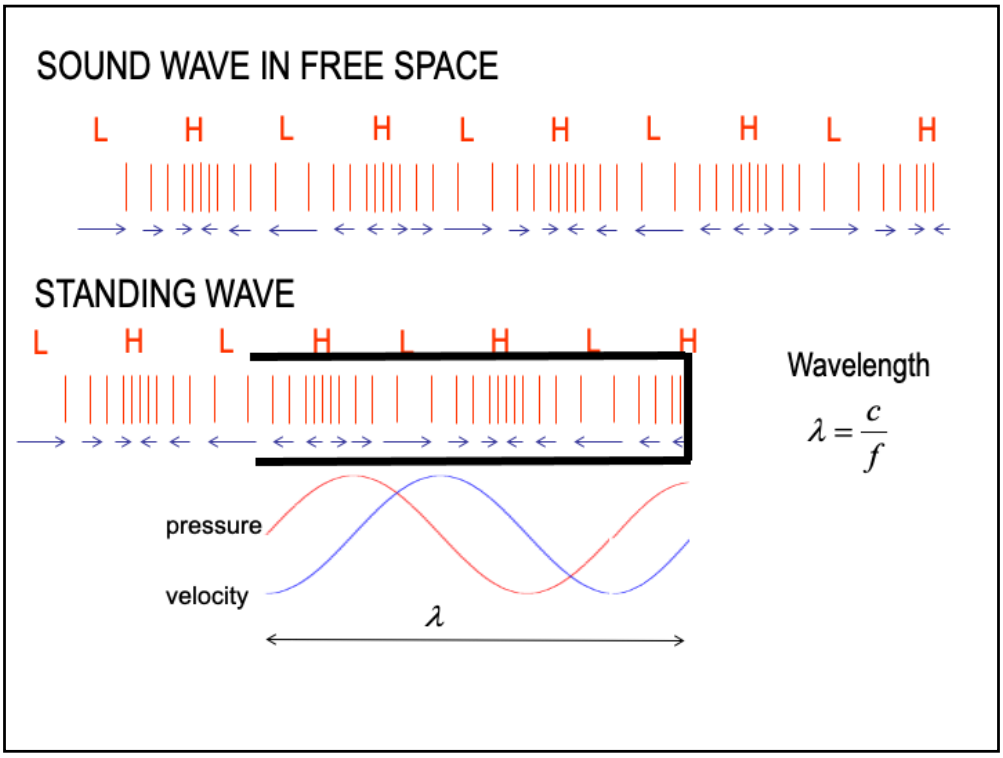
Isaac Newton

The filling of a very deeepe flaggon wth a constant streame of beere or water sounds y^{e} ^{v}owells in this order w, u, ω, o, a, e, l, y,

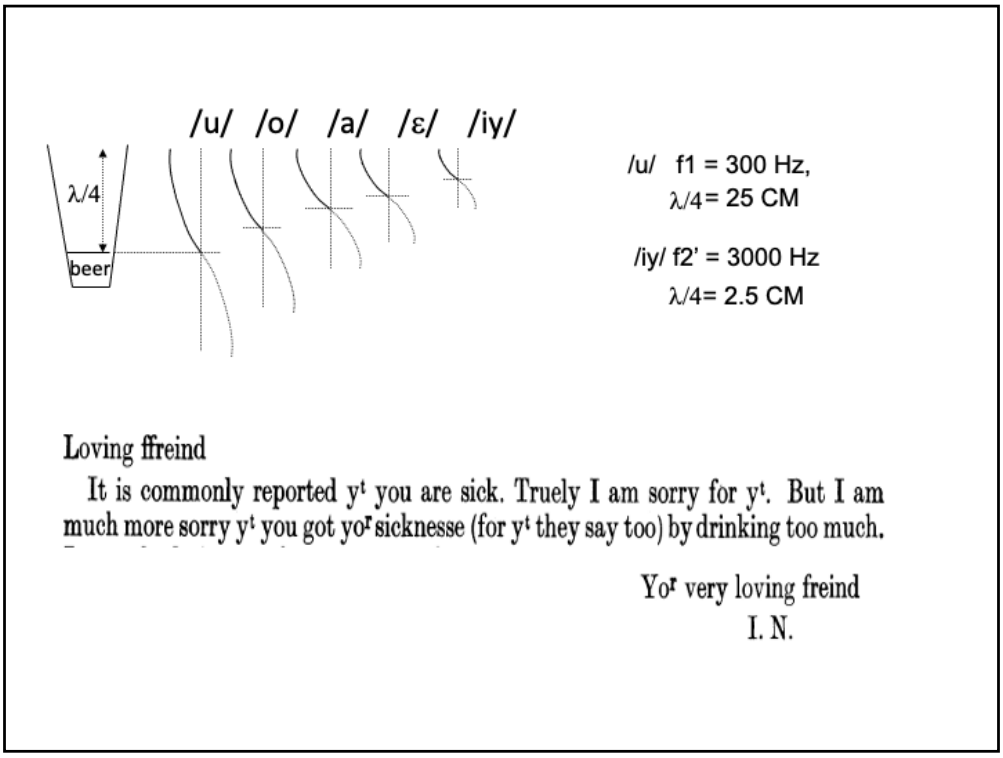
from his notes, probably 1659-1662
(17-20 years old)

Ralph W. V. Elliott: Isaac Newton as Phonetician, The Modern Language Review, Vol. 49, No. 1 (Jan., 1954), pp. 5-12

Newton tried to break different entities into their more basic elements. Here he tried to break speech vowels into simple pure tones. The device he used was a tall beer glass.

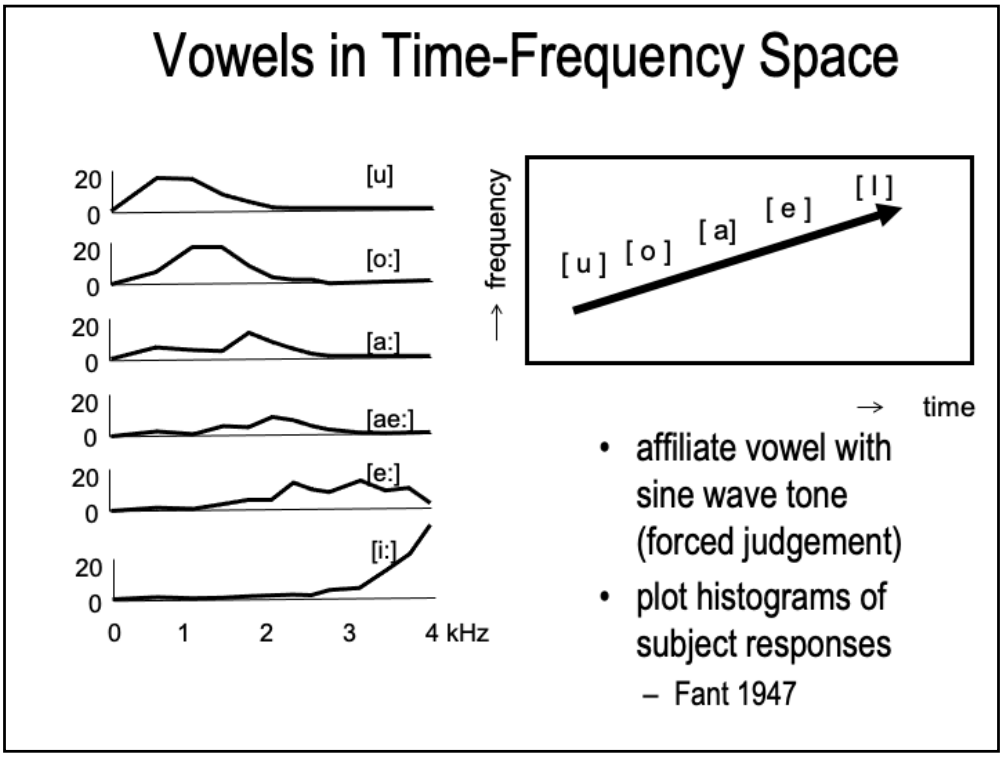


JUst to reming you, what Newton was hearing was the resonance frequency of a simple quarter-wave resonator.

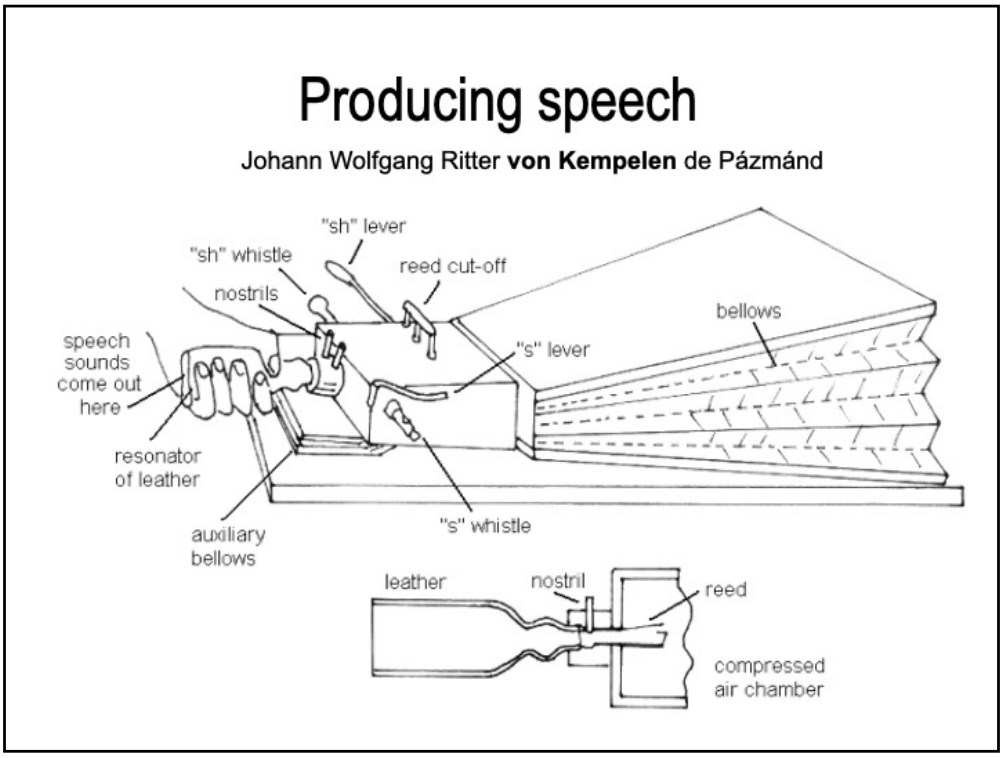


A short calculation indicated that to get /u/ he needed his glass to be at least 25 cm high. This probably involved a lot of beer drinking, since to start with the vowel /u/ the glass needed to be entirely full and since Newton did not drink at all, he must have used his friend.

Newton's appologies to his friend after the experiments were also recorded and are shown here.



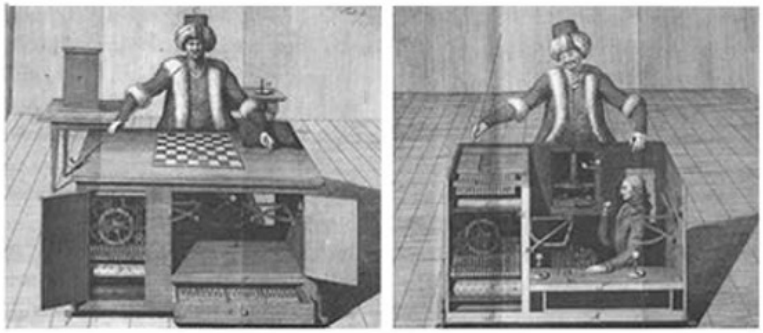
Newton's conclusion must have been that when we need to emulate whole series of voweld by simple tones, we need to go in frequencu from lowest to highest frequencies. This has been verified by some 2300 years later by Gunnar Fant ion Stockhold using forced maching of vowels by simple tones from a tone generator.



Historically, one of the first machines which emulated human speech production was mechanical synthesize of Hungarian (living in today's Bratislave-Slovakia) von Kempelen. It was very reasonable machine. Frequency rich sound was generated by vibrating reads and its spectrum was modifies by leather tube which was shaped by the hand of the operatos, just as it is modified by changes in the shape of human vocal tract.

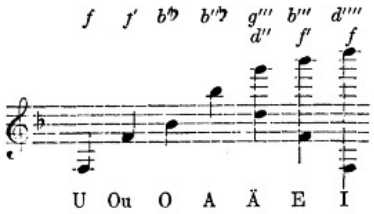
Mechanical Turk

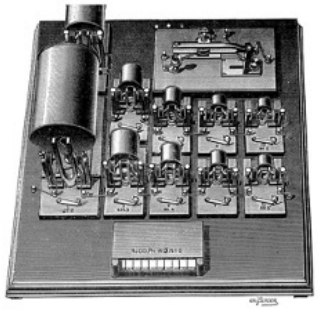
Johann Wolfgang Ritter **von Kempelen** de Pázmánd



The problem wasthat von Kempelend was a bit shady figure. Not only that he was not noble (as trying to imply by Ritter von..) but he also used his considerable intelligence for building other interesting device, such as the chess-playing Mechanical Turk (playing abd often defeating among others Napoleon Bonaparte and Benjamin Franklin) which had a skilled chess-player infide the machine ☺

von Helmholtz





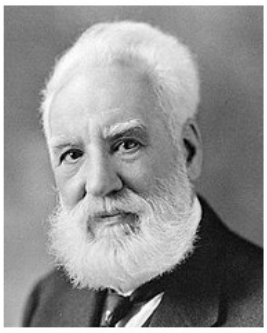
Helmholtz electric synthesizer

Newton was not the only physicist who trued to fing basic element of vowels. Von Helholtz even gives us musical notation for different vowels (he needed one tone for back vowles but prefered two tones

for the front vowels). Then it was only a step to Helmoholtz's remotely controlled music synthesizer, where he showed that **music can be transmitted on distance.**

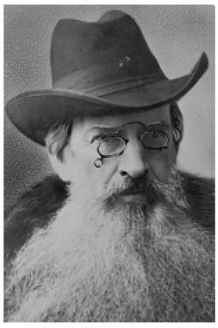
Love is all you need …..

Alexander Graham Bell



Mabel Hubbart

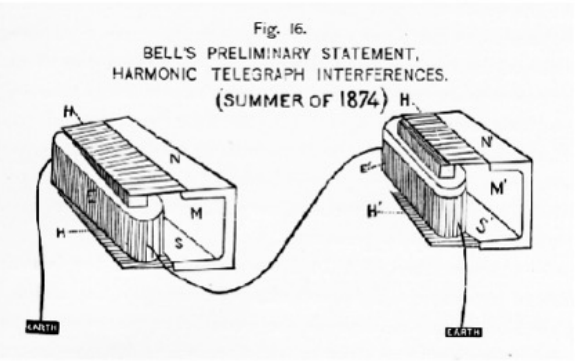
Gardiner Greene Hubbard



Vin Helmholtz/s experiments with remote music synthesis were known to the first actor of our love story, Alxander Graham Bell. He must have hoped that if one can remotely control the music syntesizer, breaking speech into the individual frequency componets (that is doing Fourier analysis) and restructuring it on the other end must be possible. However, love to his deaf student Mabel Hubbart slowed Bell's progres. Mabel's father, lawyer and money-man Gardined Humbbard asked Bell that if he wants Mabel as his wife, he should work on something more practical than telephone.

Alexander Graham Bell

Emulate spectral analysis done by hearing



Until one of the reeds got stuck and transmitted signal waveform …

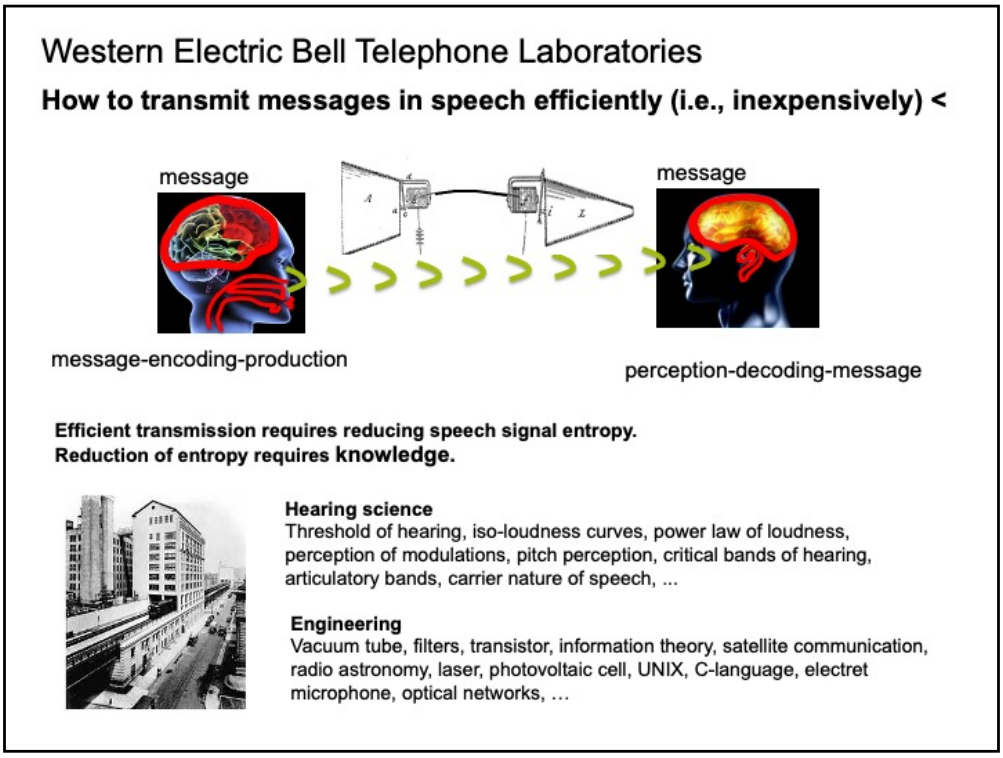
So as many other leading inventors of his times, Bell started to work on harmonic telegraph, which used principles of spectral analysis to attempt to transmit more telegraph messages on a single wire

(in today's term on frequency multiplexing). Of course Bell could not stop thinking about tranmitting voice (after all, he as his father were primarily phoneticians and both knew a lot about speech).

He could separate different frequency componets of the multiple telegraph messages and reconstruct then in the receiver using mechanical resonators shown here.

Accidetaly, while trying to fix one of the reeds on the sending side, his assitant have heard the click of of the released reed on the other side. The spectral analysis was (maybe unfortunatelly) forgotten and Bell

started to think about how to make the electric signal to change as change the acosutic pressure at speaker's mouth.



The American Bell telephone company aim was to make as much money as possible on voice telephony and they need to transmit speech as cheaply as possible.
Very quickly they started Bell Laboratories which generated many new engineering advances used in the telephony.

They knew well that so far they only managed to emulate a very small part of the human speech communication chain, the transmission of changes in acoustic pressure while speaking.

They also knew there is more to human speech communication than only speech sound.

For the efficient transmission of voice, which contains a lot of redundant elements, the reductipn of this redundancy was one of the goals and

one cannot efficiently reduce the signal redundancy without knowing what one is doing

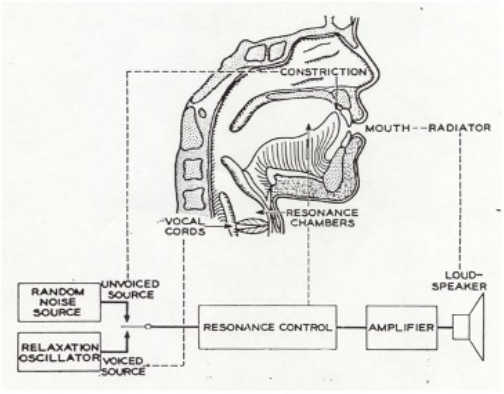
So important (if not the most important) goal Bell Labs was understanding the whole communication chain.

Subsequently, a lot of speech and hearing knowledge we use tday came from Bell Labs.

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

H. Dudley: The Carrier Nature of Speech, BSTJ 1940

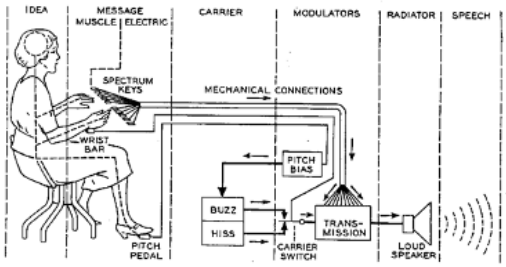
Homer Dudley



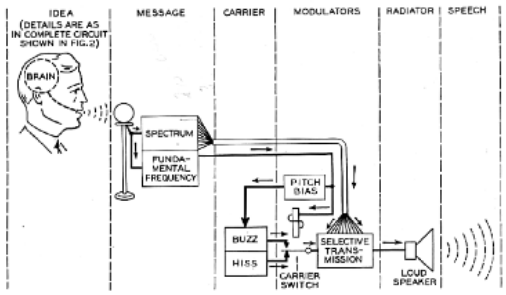
One important concept was the concept of carrier nature of speech. It says that the bulk of the information about the message in speech is in **changes of vocal tract shape,** i.e. in slow movements of

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.

Speech – message coded in muscle movements made audible by modulating the career (vocal source) by vocal tract movements



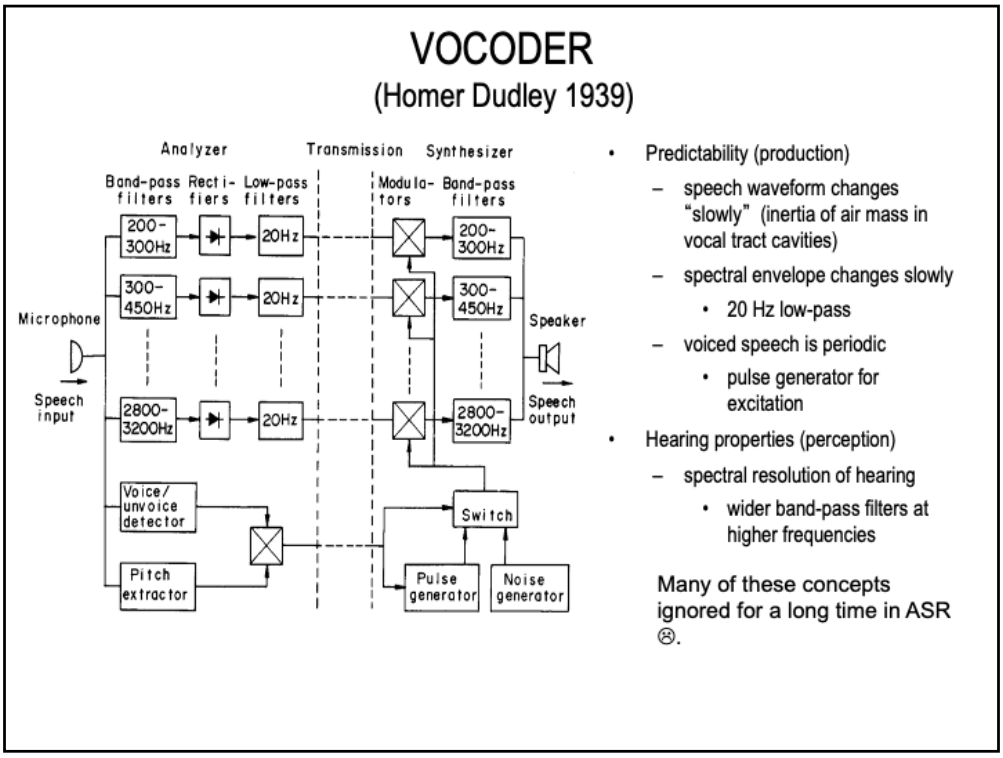
VODER
message coded in movements of fingers and legs



VOCODER
message coded in movements of vocal tract muscles

Dudley first proposed a way of nenerating human voice using a specially trained operator (she took a year to learn to operate it), who could control by hands spectral properties of the filter which shaped speech spectral envelope

and by feet the voice source which provided rich acoustic excitation signal. This snthesized was demonstrated in the Chicago World Fair in 1939. Another step from the VODER where the message was created by movements of fingers and legs, was to elminate the operator and substutute it by speech analyzer, where the message was derived from the actual movements of the vocal tract.



It is worthwihel to spenda a bit of time on the VOCODER since the VOCODER contaied many advanced concepts.

The tract movements were derived from slow (up to 20 Hz) spectral changes in freuency bands, which were gradually increasing in bandwidth with frequency as they do in human hearing accoriding to the critical band concept (also discovered in Bell labs by Harvey Fletcher) . Many of these concepts from hearing were subsequently ignored for quite a long time by ASR researchers.

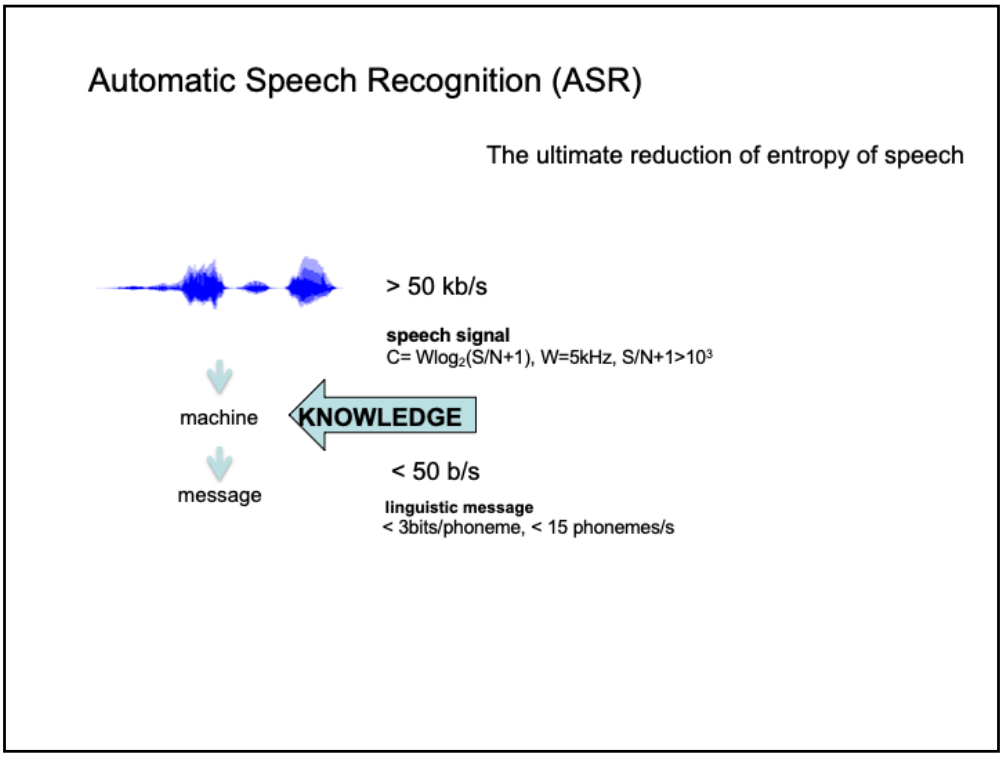
It its search for a way to counter the new submarine threat, the U.S. Navy turned to AT&T, which had been doing pioneering work with a **sound spectrograph**. Some believed this device could be used to distinguish between sounds made by submarines and background ocean noise. If so, that would allow classes of submarines to be identified by their inmistakable "acoustic fingerprints".

Mike H. Rindskopf: Steel Boats, Iron Men: History of U.S. Submarine Force.

A critical device, also developed primarily at Bell Labs, was the spectrograph. It was developed to make sounds more visible by creating 2-D time-spectral picture of acoustic signal by

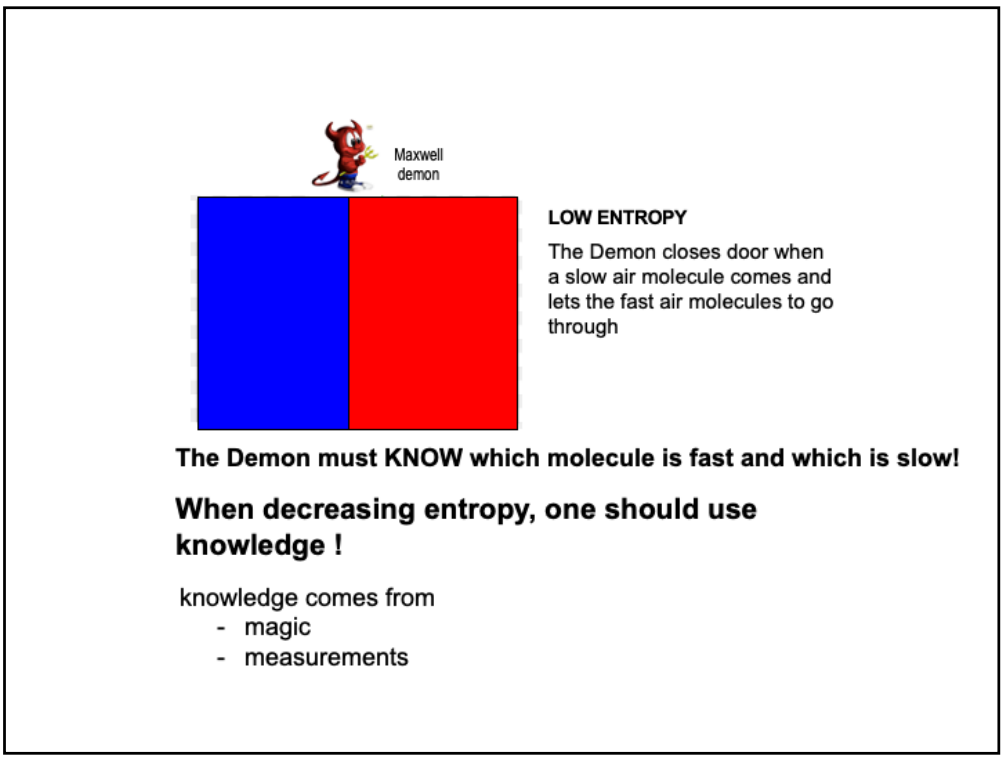
tracing energies in the individual frequency bands of the signal and was used during the 2nd WW to display underwater sounds in a hunts for enemy submarines.

At that time, it was well understood that ear is doinf frequency analysis.



The time was set for the attach on the ultimate redundancy reduction of information if messages in speech, for the more than three-orders-of-magintode reduction in speech recogntion.

Clearly, to reducethe redundancy that much, a lot of knowledge is required.



A story of the Maxwell demon may help to realize that information rate reduction is a tricky entreprise. Scotish physicist James Clerk Maxwell in 1967 pointed out that an air at a certain temperature in a closed vessel consists of slower molecultes and of faster molecules. The slow molecules make the air cooler, the fast ones make the air warmer. The particula mix of the slower and faster molecules make the air in the vessel to be of a particular temperature. Imagine that there is a partition in the vesel with a door controlled by a demon who open the door when the fact molecile comes cloe and closes it when a slow molecule comes. After a while the slow molecules are in the left part of the vessel and the fast ones are in the right part of the vessel. The vessel became better organized, we know more about probabilities of molecules speeds in both parts of the vessel, and the entropy of the decreared, The temperatore gradient was created which could be used for creating an energy! The second law of thermodynamics was violated and a *perpetum mobile* was created.

Many tried to explain wher is the error in thinking. An interesting explanation is one from Leo Szilárd who poitde out that the demon needs **knowledge** to do its task, acquired by measurements which requires some energy. The entropy reduction of the gass requires increase of the knowledge (entropy) of the demon, so the whole system stays at the same entropy. Entropy decrease requires knowledge increase. The whole argument still goes on. We can now

even know how much energy is in one bit of information [Landauer 1961].

For our purpose, we should remember that reduction of bit rate (entropy) in speech by recognizing the string of sounds which form the message requires knowledge.

KNOWLEDGE

- magic
- experts, beliefs, previous experience (hardwired)
- measurements (data)

HARDWIRED

– experts and beliefs can be wrong

but

– no need to re-learn known facts

DATA

– transcribed data are expensive

but

– data do not lie

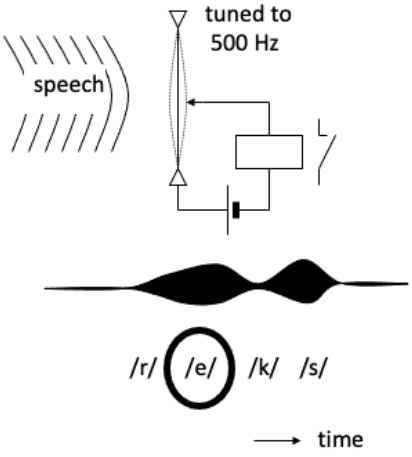
Apart of the magic, the available knowledge can be hardwired into a machine or to start with a clean machine and to acquired the necessary knowledge from training data.

As for the hardwired knowledge, it was hopefully also acquired earlier from the data and it might have been formulized in textbooks. Its proponts are correct in saying that what is already known, does not need to be learned again.

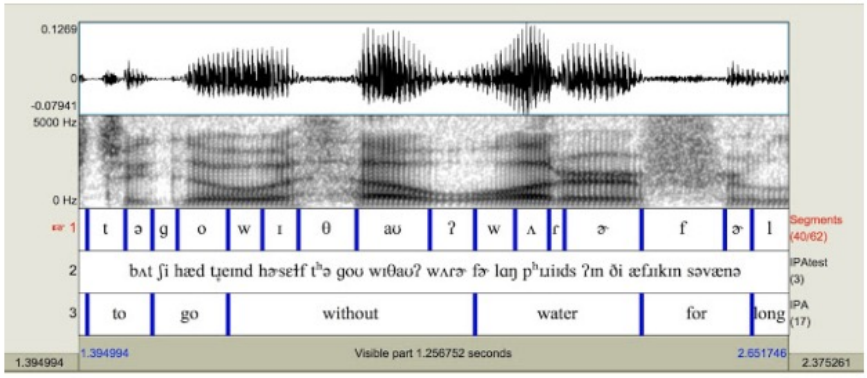
In that way, the need for amounts of training data may be reduced. The problem might be that wrong or incomplete hardwired knowledge may be harmful.

For the *tabula rasa* approach, the large amounts of transcribed training data may be expensive to acquire. However, no prior constraints may lead to better results since only the knowledge which is relevant to speech recogntion is being used.

Radio Rex
The first "knowledge-based" speech recognizer



When it comes to speech recogntions we need to know the first succesful commercial product. The dog house forms an acoustic resonator (remember our speech production lecture) which resonates around 500 Hz where the vowel /e/ in Rex has concentration of spectral energy. The house walls resonate when the name of the Dox is called, a small contact on the side of the house disconnets, electromagnet stopd holding the spring and the dox ois pushed out. The machine was first patented in 1913.



courtessy of blogger “Buffalo LInguist”
http://christiandicanio.blogspot.com/2018/12/what-is-phonetics-20-minute-guide-for.html

Spectrograms are just too tempting as a source of knowledge. WIth some (considerable) training, one can learn how to read spectrgram of single adult speakers, so one can make a case that there is enoufg information to decode the speech messages from such spectral represnations. In spite of the early failure of speech recogntion based on spectrograms, the short-time spectra and the associated spectrograms still in some forms with us.

WW-2 and German submarines

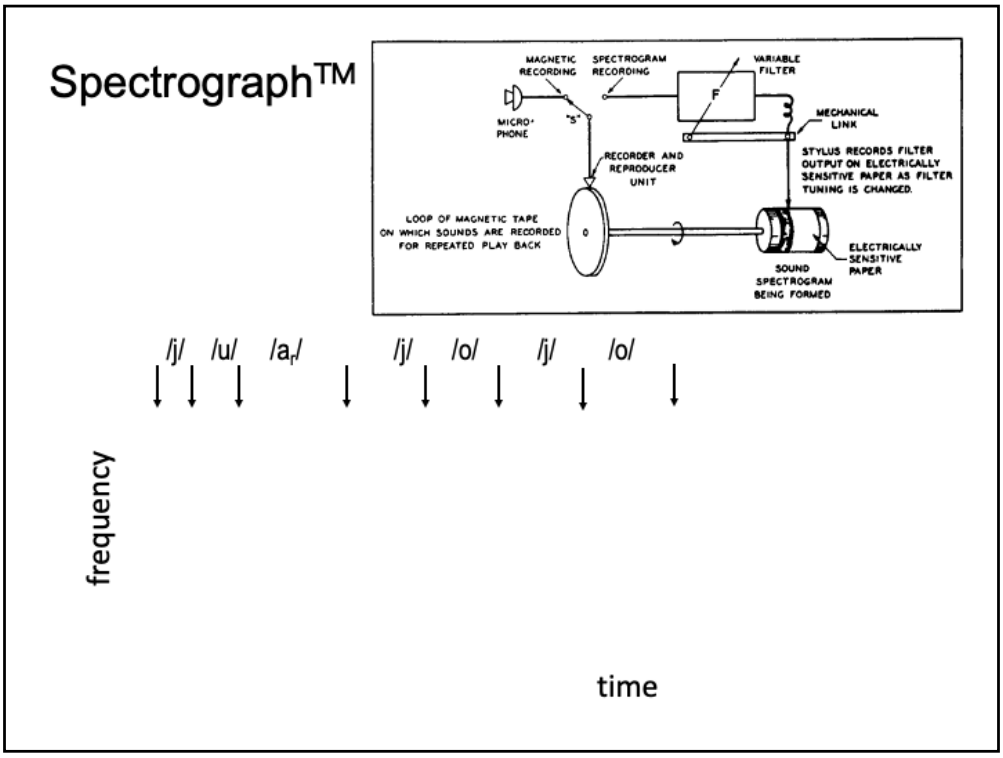
It its search for a way to counter the new submarine threat, the U.S. Navy turned to AT&T, which had been doing pioneering work with a **sound spectrograph**. Some believed this device could be used to distinguish between sounds made by submarines and background ocean noise. If so, that would allow classes of submarines to be identified by their inmistakable "acoustic fingerprints".

Mike H. Rindskopf: Steel Boats, Iron Men: History of U.S. Submarine Force.

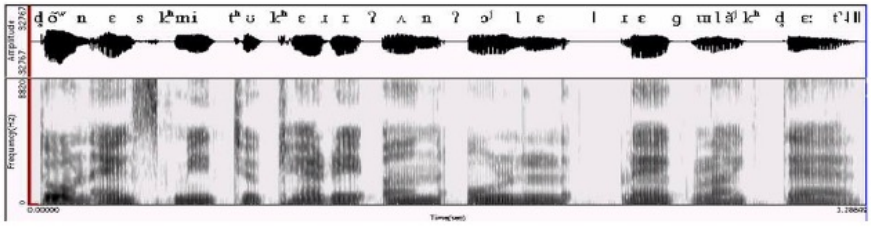
A critical device, also developed primarily at Bell Labs, was the spectrograph. It was developed to make sounds more visible by creating 2-D time-spectral picture of acoustic signal by

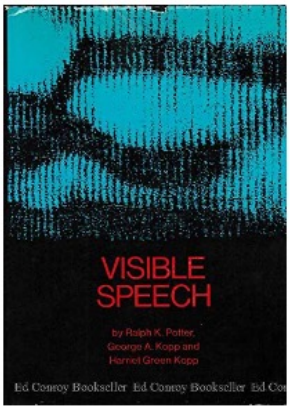
tracing energies in the individual frequency bands of the signal and was used during the 2nd WW to display underwater sounds in a hunts for enemy submarines.

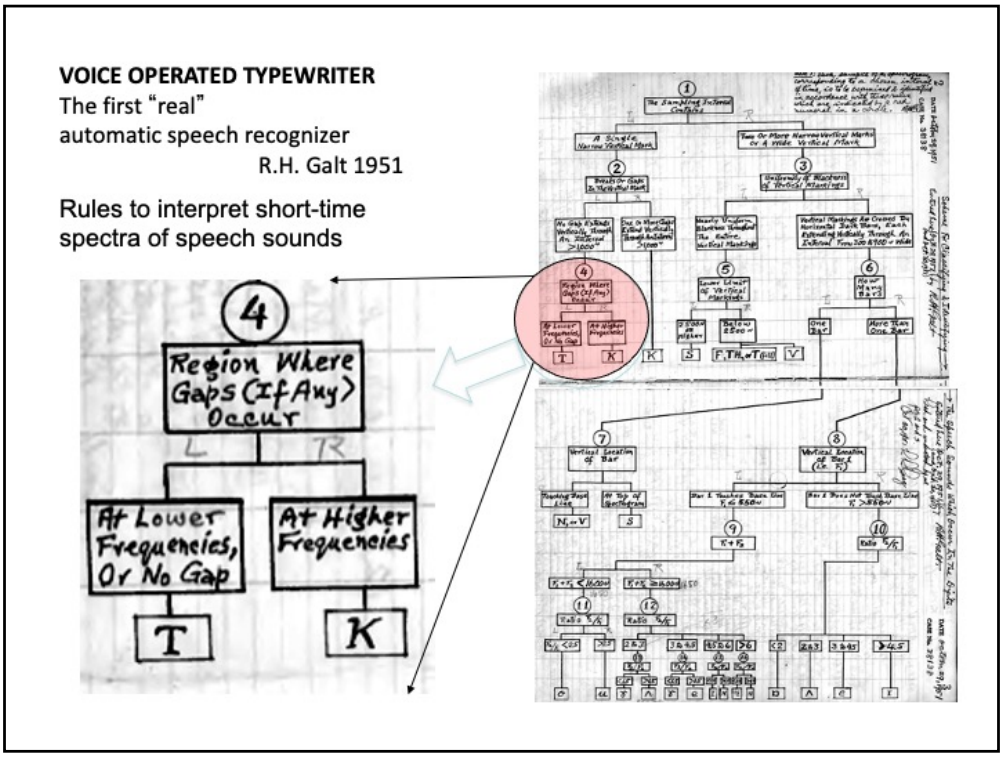
At that time, it was well understood that ear is doinf frequency analysis.



Here we see how the mostly mechanical Spectrograph™ was working. The signal was recorded on magnetic disk and played back trough a bandpass filter which was toned to different pass band frequencies as the spectral power at the filter output was recorsde as a finction of time and the passband frequency on a paper.



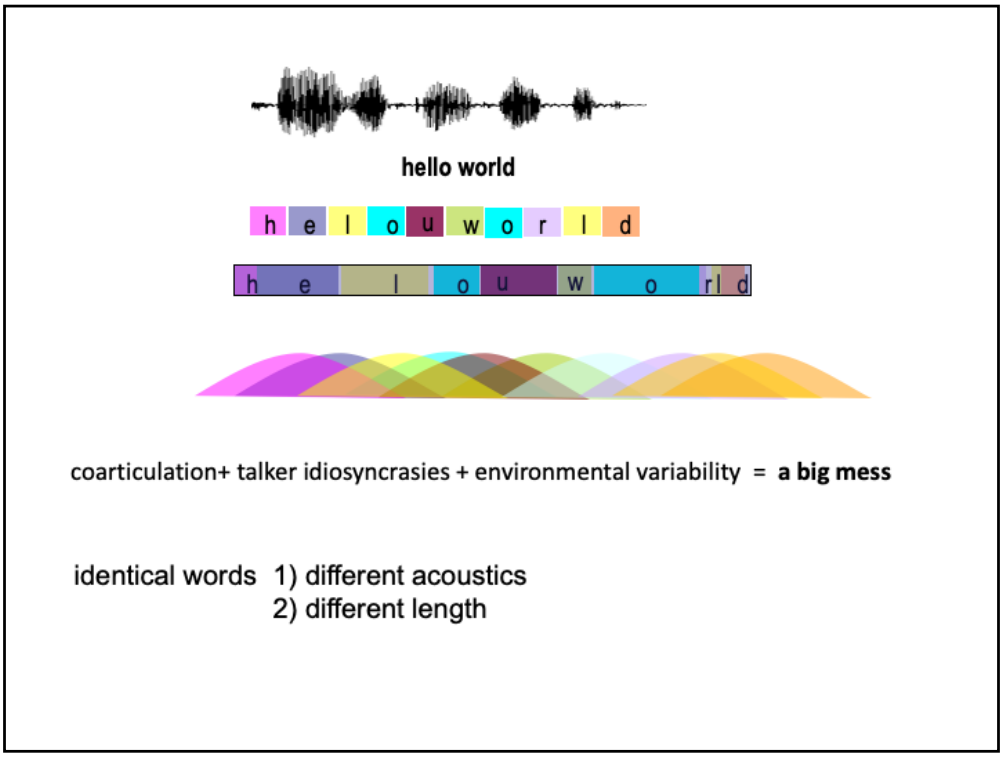

VISIBLE
SPEECH
by Ralph K. Potter,
George A. Kopp and
Harriet Green Kopp
Ed Conroy Bookseller Ed Conroy Bookseller



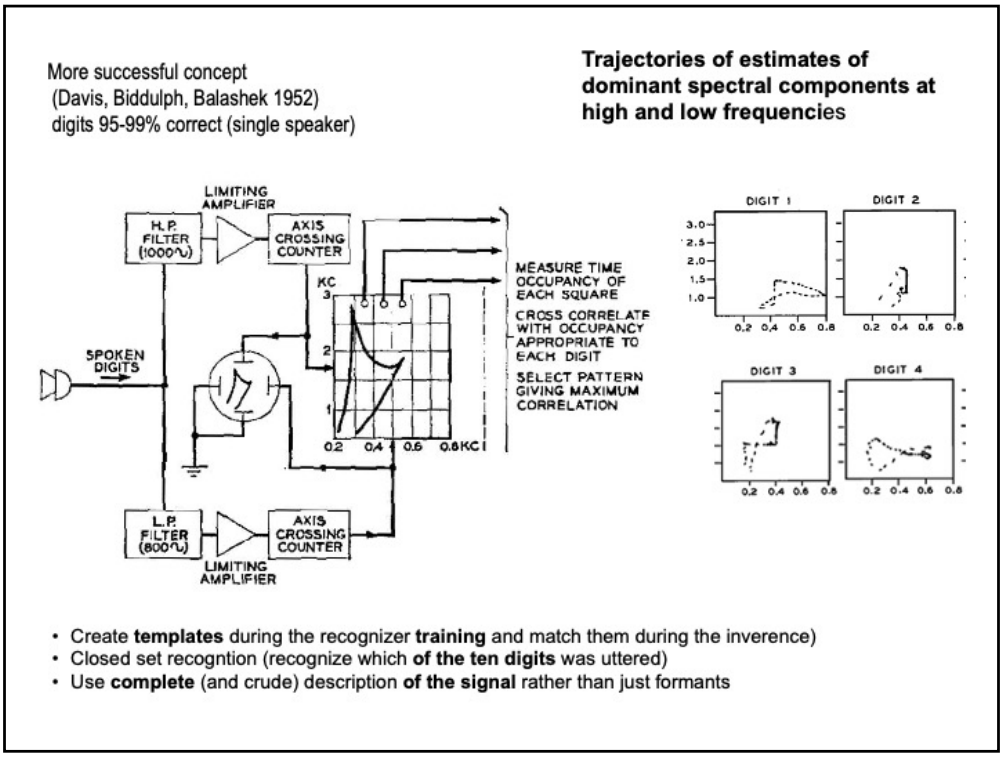
.

R. Galt of Bell Labs collected a lot of knowledge how his colleagues read the spectrograms and attempted to build

the first decision-tree based "knowledge-based" recognizer. One can only admire the maouhnt of work put into this design.



Coarticulations modify acoustics of speech sounds by influences from neigbouring speech sounds. In addition, speakers can speak faster or slower and modify relative lenghts of different speech segments while speech messages stay the same.



Competing group at Bell Labs took different approach. The only “knowledge” applied here was the fact that human hearing is analyzing sounds onto its spectral componets (remember Newton?)

and crudely measured main clusters of energies (zero-crossings) in the low part and the high part of speech spectrum. Displaying paterns of these measurements for each of ten digits they traied to recognizer

on the 2-D plane, creating the digit **templates.** Using rather ingenious electronics (not shown here) for remembering the templates and for matching during the inference,

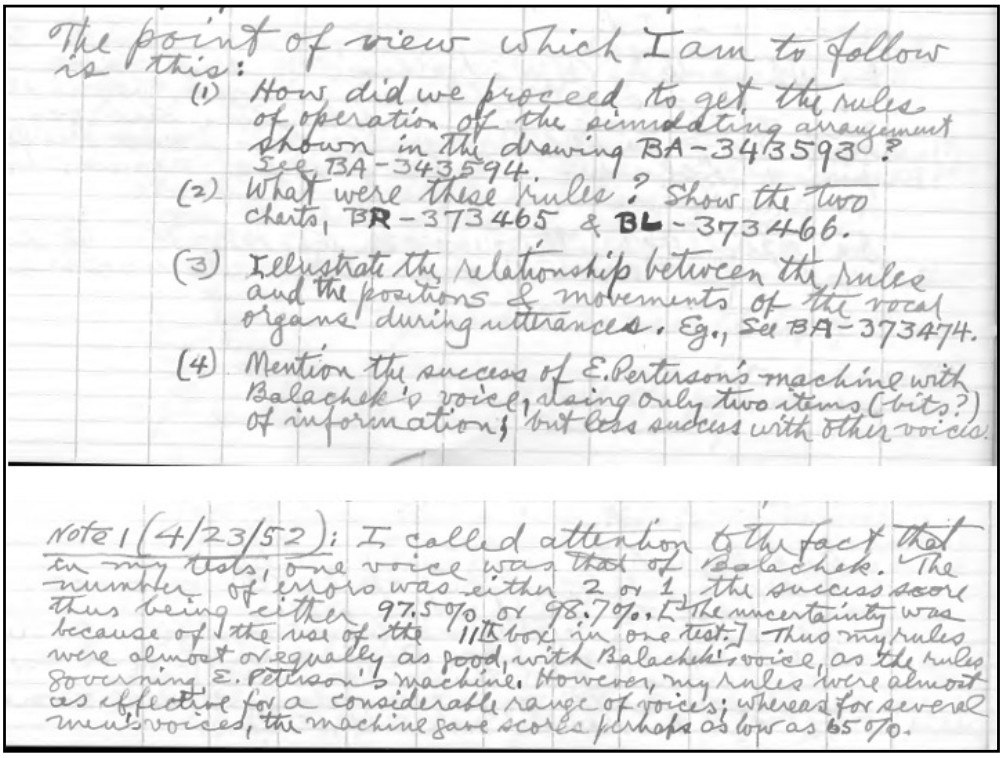
the recognizer achieved > 98% accuracy on a single speaker for whom it was trained.

Several concepts, which in some forms carry over all the way to the current ASR technologies were introduced.

•Template matching (create templates during the recognizer training and match them during the inverence)

•Closed set recogntion (recognize which of the ten digits was uttered, do not allow for extraneous sounds)

•Use complete (although crude) description of the signal rather than just formants



The point of view which I am to follow is this:

(1) How did we proceed to get the rules of operation of the simulating arrangement shown in the drawing BA-343593? See BA-343594.

(2) What were these rules? Show the two charts, BR-373465 & BL-373466.

(3) Illustrate the relationship between the rules and the positions & movements of the vocal organs during utterances. Eg., See BA-373474.

(4) Mention the success of E. Peterson's machine with Balachek's voice, using only two items (bits?) of information; but less success with other voices.

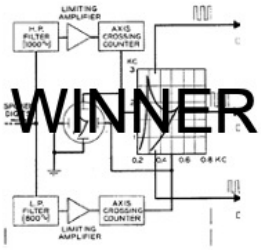
Note 1 (4/23/52): I called attention to the fact that in my tests, one voice was that of Balachek. The number of errors was either 2 or 1, the success score thus being either 97.5% or 98.7%. The uncertainty was because of the use of the "11" box in one test. Thus my rules were almost or equally as good, with Balachek's voice, as the rules governing E. Peterson's machine. However, my rules were almost as effective for a considerable range of voices; whereas for several men's voices, the machine gave scores perhaps as low as 65%.

To give you some sense of the drama surrounding the demos of the first speech recognizers at Bell Labs, here are copies of few pages from Dr. Galt's notebook (all Bell Labs engineeres were required to keep such notebooks. describing their intermediate results and thoughts).

The point which Dr. Galt corretly made was that his techniques was realtively speaker independent and capable to recognize more that the ten digits of the compeeting solution of Davis, Biddulph and Balasek. . However, the competing solution prevail;es and Mr. Galt was put on another project. Even the liberal Bell Labs did not allow complete research freedom and kept their researchers focused on promissing approaches.

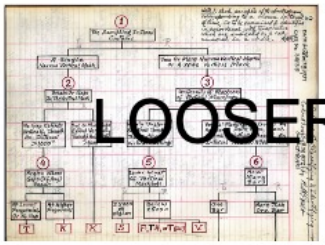
TASK: Recognize closed set of ten digits (important for Bell System)

DATA: DIGIT RECOGNIZER



Close set recogntion
Speaker-specific
High and low frequency signal energy
Word templates

RULES: VOICE OPERATED TYPEWRITER



Open vocabulary
Speaker-independent
Spectrogram-based (formant patterns)
Phoneme string spotting

good:
build tools (digit recognizer) and not artificial human listeners
what you see (spectrogram) is not what we listen for (spectral dynamics)

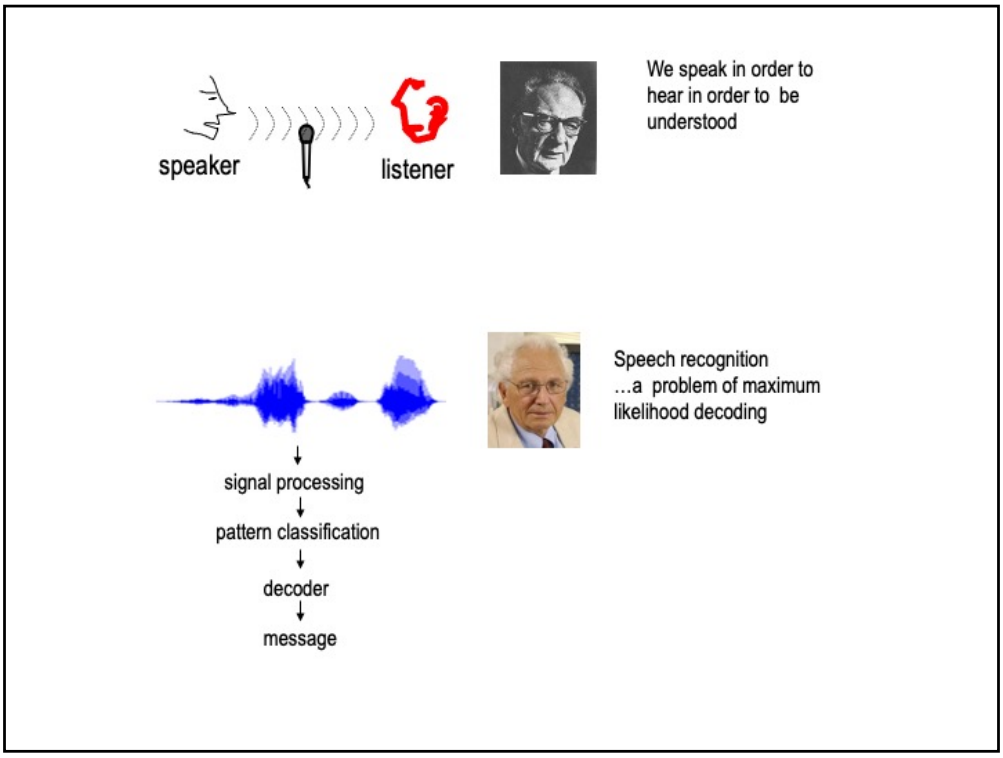
maybe bad:
Implicit assumption that training data represent the present situation
focus on pattern matching (expected) and not on information (unexpected)

No surprise, the template matching concept was the winner (it was eventually even produced in small numbers under the name of Audrey).

Important lessons were learned. Some of the concepts applied in the winning system carry over to current ASR.

One can only wonder if the task Bell Labs set up first was to **recognize a single word (keyword spotting)** in the background of all possible sounds,

This would have been more relevant to human speech communication and our recognizers could have looked different.



The two approaches can be summarized as here. The knowledge to be hardwired should be related to our knowledge about human speech coimmnunication process. The knowledge to be extracted from the data should aim dirrectly for the fulfilling the task of extraction information from signal by machine (machine learning).

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

Short-time spectra which form bases of spectrogram are fragile. Even though carefully articulated speech produced in quiet environments can deliver amaizingly clear spectrograms, very minor disturbances which are hardly noticed by human listeners can modift the spectrogram considerably.

Template matching

Bell Labs 1952

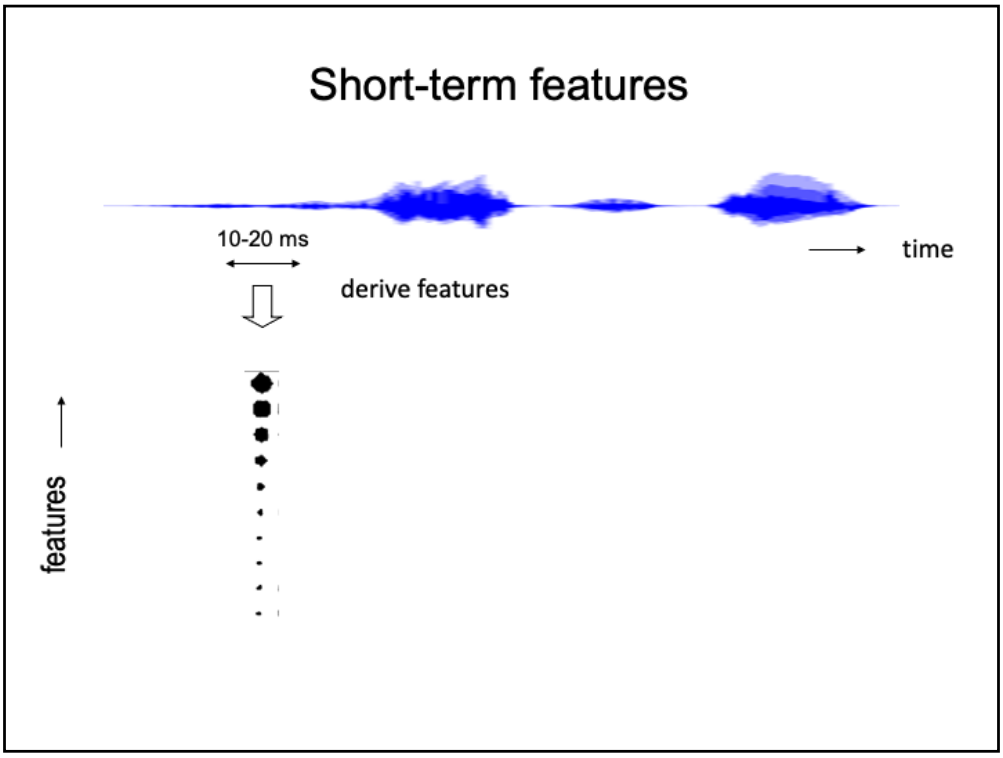
Compare sequences 2-D parameters

1. Derive examples of expected utterances from training data
2. Find the best match of your test data with the previously derived examples

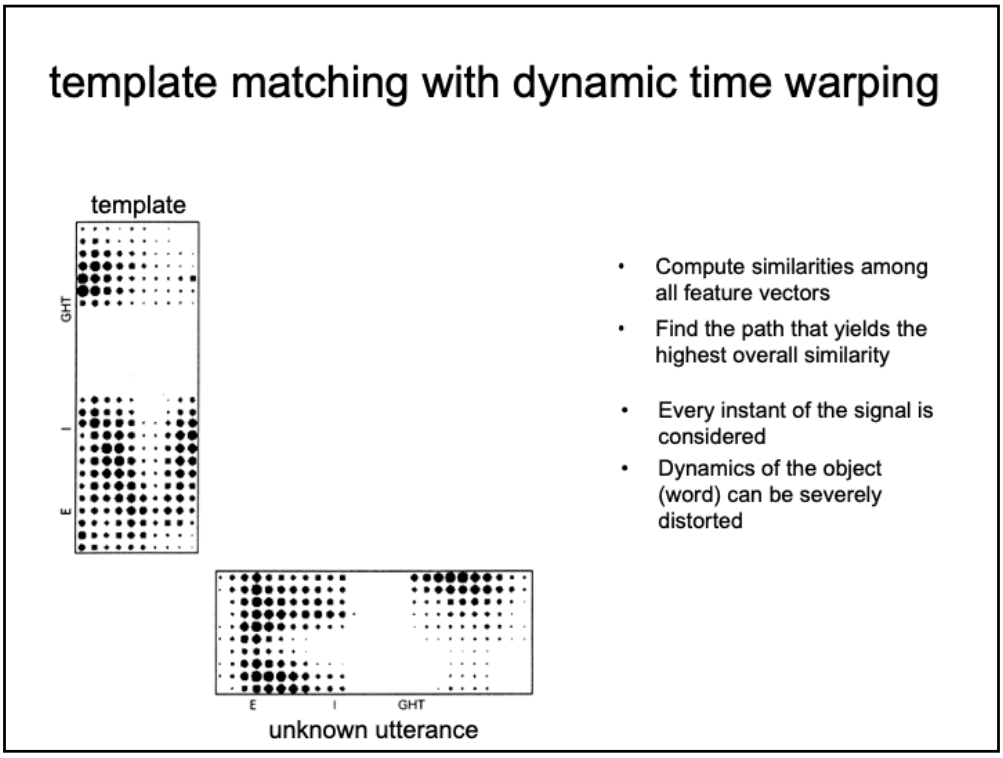
Bell Labs (Fumidata Itakura from NTT) 1974

Compare sequences of muti-dimensional parameters

However, to deal with differences in timings, more powerful soulution is offered by dynamic time warping, introduced in seventies independently in Japan and in Soviet Union

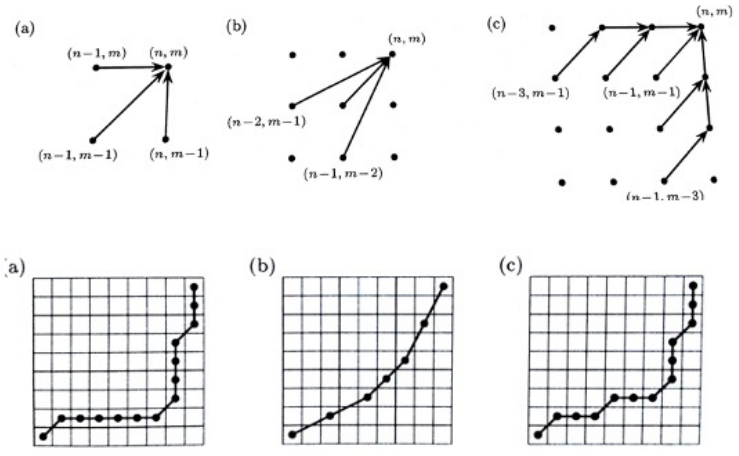


To get more information, we may want to move to more dimensions but the two provided by the high and low zerocrossing frequencies.

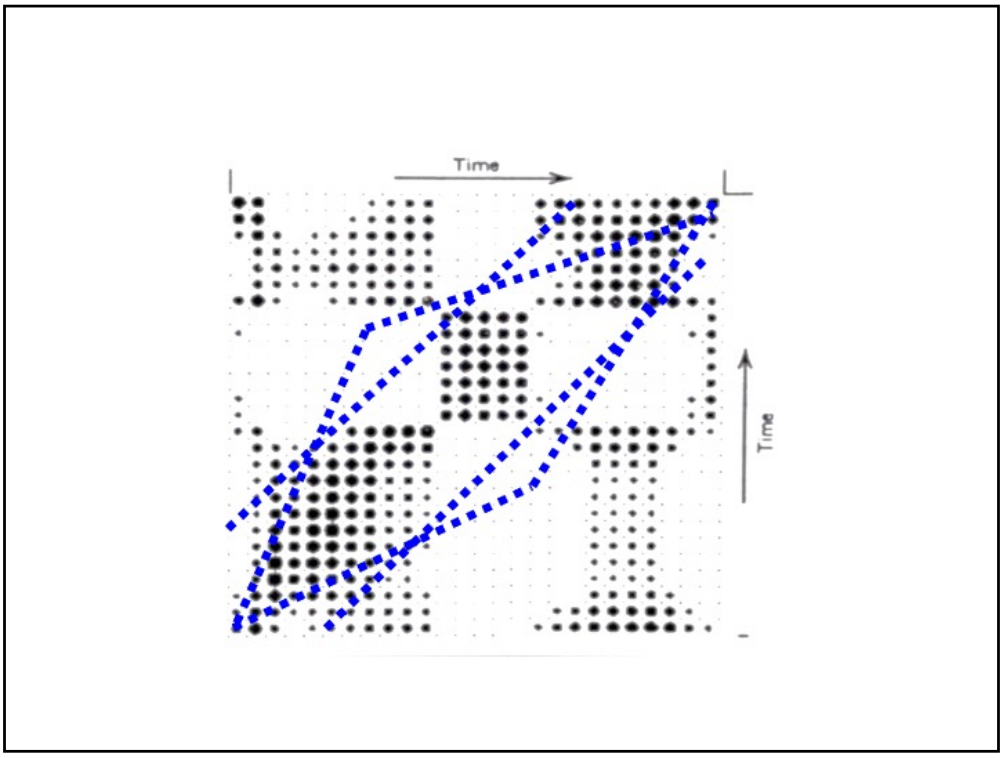


To account for the fact that different people may have different speaking habits and the utterances with identical linguistic messages may have different timing, the dunamic time warting technique was introduces to speech recogntion. The two utterances are compared so that the minimum distance (the maximum similarity) between them is achieved while allowing for local adjustments of the path along which the utterances are compared. MATLAB program from the signal processing toolbox dist = dtw(x,y) stretches two vectors, x and y, onto a common set of instants such that dist, the sum of the Euclidean distances between corresponding points, is smallest. To stretch the inputs, dtw repeats each element of x and y as many times as necessary. If x and y are matrices, then dist stretches them by repeating their columns. In that case, x and y must have the same number of rows.

How to form the “best” path

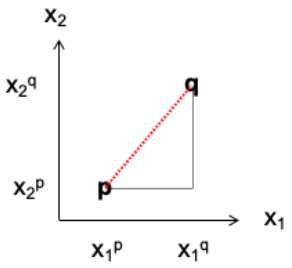


Depending on a particular technique to do the local time warping, different paths are taken to find the best match. To reach the next point (m,n), here are at least three ways to do so.



Additionally, to prevent impossible alignments (and to save on computations) the area which the path can take is typically limited. Here are two ways of doing so but your fantasy can also come with different ones.

Distance between points in vector space



in n dimensions

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$ Euclidean

Generalization

$$d_M(p,q) = \left(\sum_{i=1}^{n} (p_i - q_i)^m\right)^{1/m}$$ Minkowski

To compute the cummulative distance, we need first to decide on an appropriate local distance computation. The simplest are shown here.

Mahanalobis distance
(assumes Normal distribution of parameters)

$$d_m = \sqrt{(\boldsymbol{p}-\boldsymbol{q})^T \Sigma^{-1} (\boldsymbol{p}-\boldsymbol{q})}$$

Σ – covariance matrix of the vector space
where $\boldsymbol{p}$ and $\boldsymbol{q}$ lays

simplifies to inverse variance weighted distance when features uncorrelated

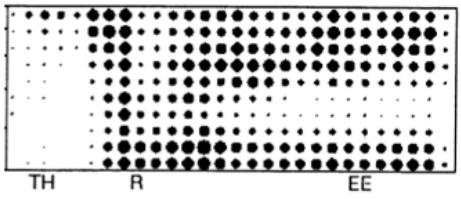
simplifies further to Euclidean distance when all features have equal variance

Euclidean distance assumes Normal distribution of parameters that are uncorrelated and have all equal variances

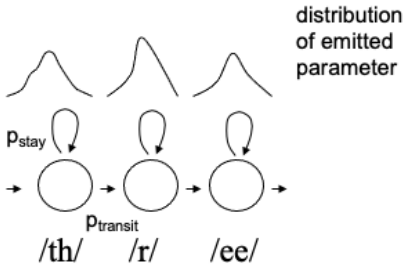
The general distance to deal with data which may have unequal variances and be uncorrelated in the Mahanalobis distance. This distance simplifies to the weighted Euclidean distance when the data are uncorrelated (de covariance matrix is diagonal) and to the Euclidean distance when the variances of the data in all dimensional are the same. (then the covariance matric is the identity matrix)

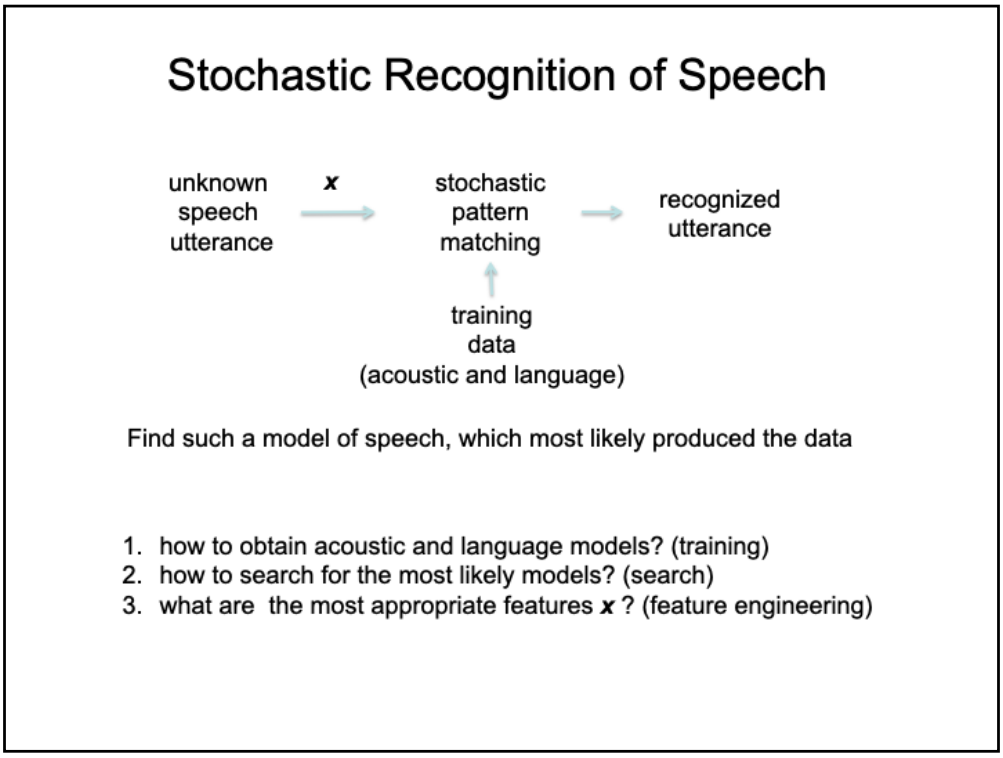
Model ?

- Examples of utterances (templates)
- similarity by evaluating "distances"



- Generative stochastic model (hidden markov)
- similarity by evaluating likelihoods of models





The stochastic recognizer finds the best model which could produce the unknown data ***x.*** The stochastic matching module needs to generate all possible uttereances which could produce the data. This is done using acoustic and language models derived from data. There are three main issues with this approach. 1) How to define the input data ***x*** (the features), 2) how to train the acoustic and the language models, 3) how to search for the most liklely models. SInce models are build from training data, the implicit assumption is that training data represent the present situation

Hidden Markov Model

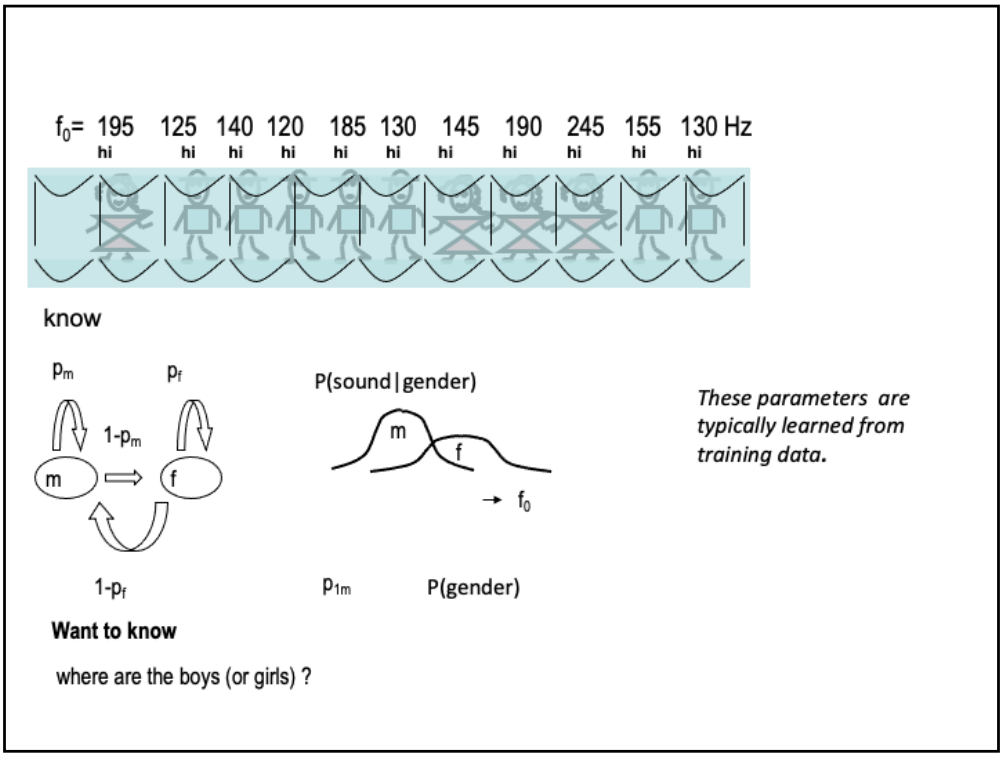
Two dominant sources of variability in speech

1. different people sound different, communication environment different,… (feature variability)
2. people say the same thing with different speeds (temporal variability)

"Doubly stochastic" process (Hidden Markov Model)

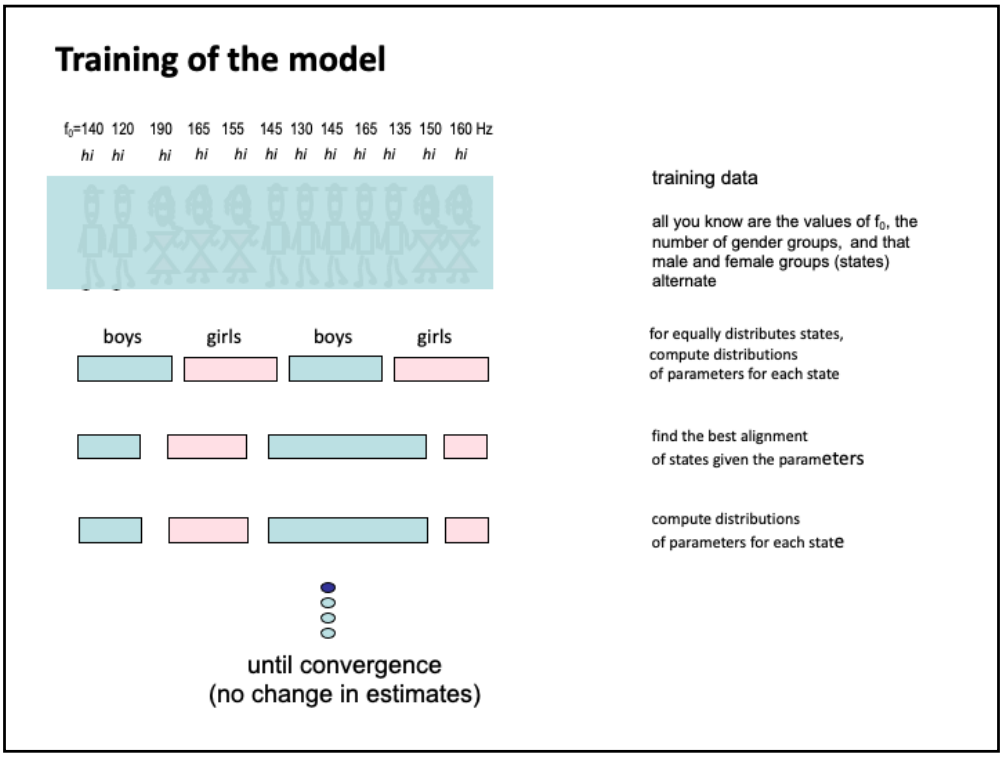
Speech as a sequence of hidden states (phonemes) - recover the sequence

1. never know for sure which data will be generated from a given state
2. never know for sure in which state we are

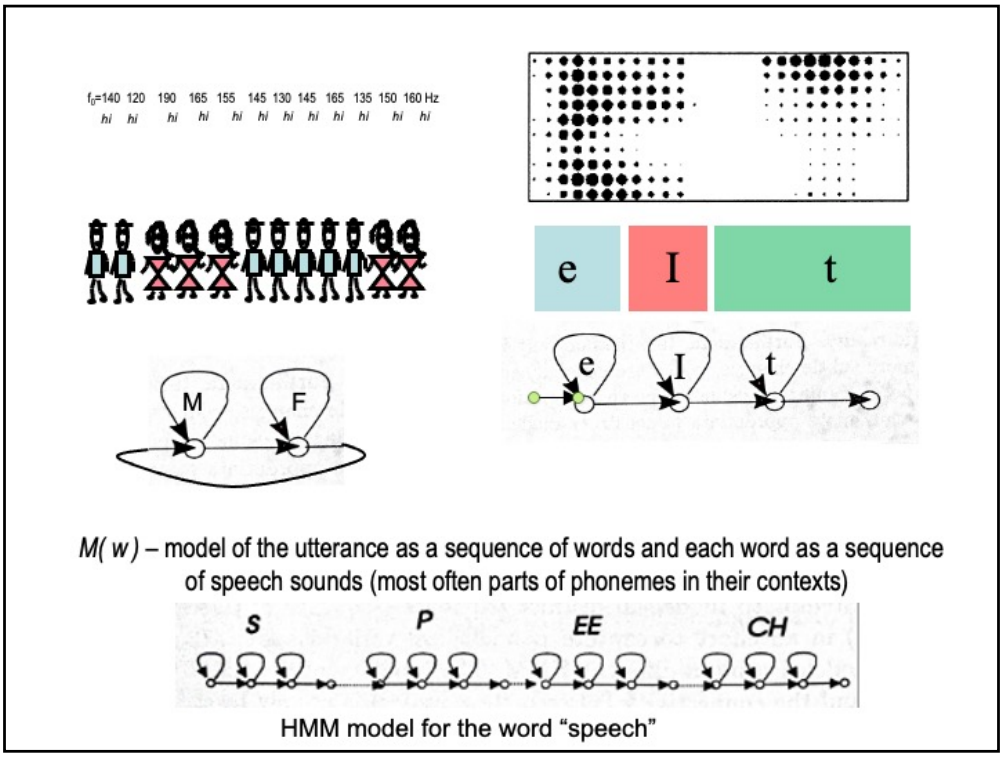


Imagine that there is a line of groups of men and women. The are hidden behind the curtain. Each person says a simple utterance. You can heasure some parameter from each utterance (e.g., the pitch of the voise of a given person).

You have a simple model where you know that when you are in a particular group, you know the probability distributions of sounds from each of the gender groups, and can either stay in the group with the probability of p_m for the male group and p_f for the female group or to transit to the oposite group. Your task is to discover where are the male and where are the female groups in this line.

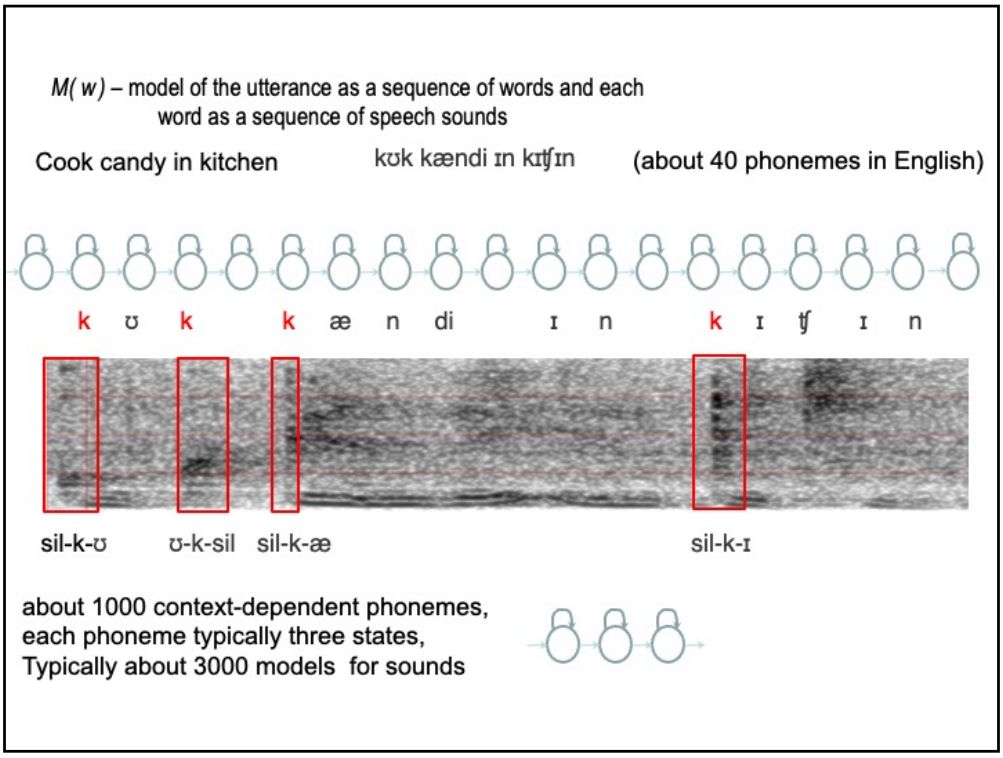


To be able to recognize the group sequences, you need the distributions of the group sounds and the transition probabilities between groups. These can be obtained from the training data. First you need to know how many groups are in your training data (transccription of the data). You do not know where the boundaries between the groups are, so you assume that the groups are distribute in the space equally. You compute means of the groups and the transition probabilities under this assumption and realign the data using such obtained model parameters. You recompute the model parameters form the new alignment and recompute the model parameters. You can continue this process until there is no change in the boundaries (or the changes in the boundaries are negligible).



Imagine that instead of sequences of people of two different genders, you have senquences of speech segments, groups of which represent different speech sounds. The principles we discussed for the groups of people now apply to groups of speech segments. Instead of a single measuremens of the pitch of "hi" we now mofe detailed descriptions of sounds Imn terms of feature vectors (for now imagine, e.g., the short-time spectral vectors). Sequences of speech sounds can be concatenated into words, sequences of words into phrases, e.t.c.

To deal with coarticulations among phonemes, the individual speech sounds which may represent phonemes of the language can be further subdivided into shorter subsegments (often three subsegments, representing the first, the middle and the last part of the phoneme). Instead of a simple cphoneme classes, we could create many more classes of so called cothert-dependent phonemes, which differ depending on their phonetic contexts. …The variations can be many and many have been tried.



Most often, classes are not just simple phonemes (around 50 of them) but rather so called “context-dependednt phonemes”, which represent phonemes in particular phoneme contexts. Her we show phoneme /k/ , the fisrt being in the conxtext of left silence and right ʊ while another one is in the context left ʊ and right silence, and the next one in the context of left silence and right ae Spectrogram shows that they are acousticaly different due to coarticulation. The context phoneme classes are typically clustered to limit the number of final phoneme classes but still the trypical recogntion systém uses several thousand of them. So the model can be getting more and more complex and requiers more and more ttraining data.

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.

Since 1969

- Better speech features (linear prediction, cepstrum, auditory-like techniques…)
- Better pattern matching (dynamic time warping, Viterbi search)
- Stochastic models allowing for using huge amounts of speech data
- Iterative expectation-maximization (EM based training only from transcribed speech data (no need for data labeling)
- Explicit use of Bayes rule combining the evidence from the signal together with prior expectations form the language

49

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.

Stochastic machine recognition of speech

$$P(M,x) = P(M|x)P(x) = P(x|M)P(M)$$

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

$$M = \operatorname{argmax} P(x|M_i)P(M_i)$$

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.