Chapter 1

Introduction to Speech Signal Processing

语音信号处理概述

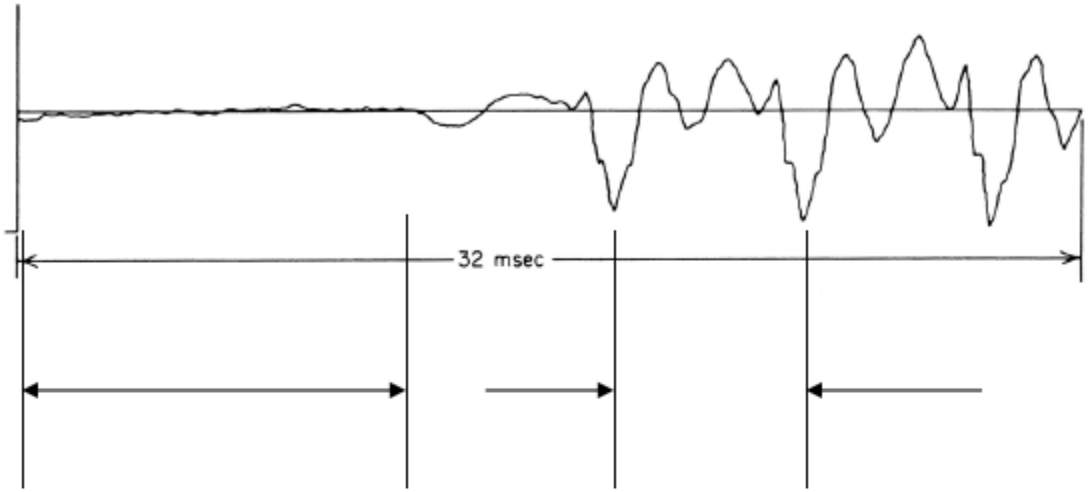
Outline

- The Speech Signal
- Speech Signal Processing
- Speech Production/Perception Model and the Speech Chain
- The Speech Stack
- Applications of Speech Signal Processing
- History of Speech Signal Processing

The Speech Signal

- Speech(语音) is the vocalized(有声的) form of human communication
- The fundamental purpose of speech is human communication; i.e., the transmission of messages(信息) between a speaker and a listener
- The fundamental analog form of the message is an acoustic waveform(声学波形) that we call the speech signal(语音信号)
- Speech signals can be
  - converted to an electrical waveform by a microphone
  - manipulated by analog/digital signal processing
  - converted back to acoustic form by a loudspeaker/headphone

The Speech Signal



Background Signal

Pitch Period

Unvoiced Signal (noise-like sound)

Software

- Praat
  - http://www.fon.hum.uva.nl/praat/

- Cool Edit Pro (Adobe Audition)

Speech Signal Processing

- Speech Signal Processing (语音信号处理)
  - converting one type of speech signal representation to another so as to uncover various mathematical or practical properties of the speech signal (发掘语音特征) and do appropriate processing to aid in solving both fundamental and deep problems of interest (解决实际问题)
- Purpose of speech signal processing
  - To understand speech as a means of communication
  - To represent speech for transmission and reproduction
  - To analyze speech for automatic recognition and extraction of information
  - To discover some physiological characteristics of the talker

Speech Signal Processing

- Digital processing of speech signal (数字语音信号处理, DPSS)
  - obtaining discrete representations of speech signal，which preserves the information content in the speech signal, also it is convenient for transmission or storage
  - theory, design and implementation of numerical procedures (algorithms) for processing the discrete representation in order to achieve a goal (recognizing the signal, modifying the time scale of the signal, removing background noise from the signal, etc.)

Speech Signal Processing

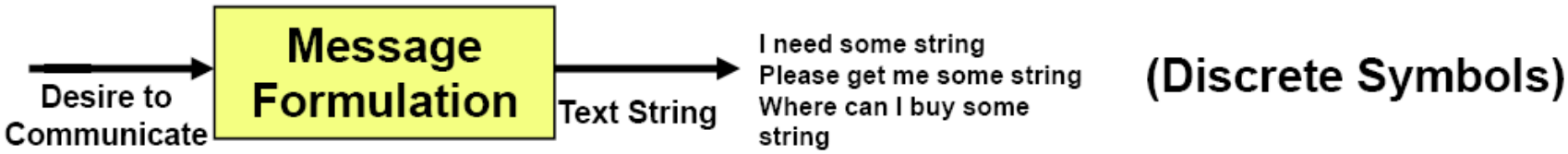
- Advantages of DPSS
  - reliability
  - flexibility
  - accuracy
  - real-time implementations on inexpensive DSP chips
  - ability to integrate with multimedia and data
  - encryptability/security of the data and the data representations via suitable techniques

Outline

- The Speech Signal
- Speech Signal Processing
- Speech Production/Perception Model and the Speech Chain
- The Speech Stack
- Applications of Speech Signal Processing
- History of Speech Signal Processing

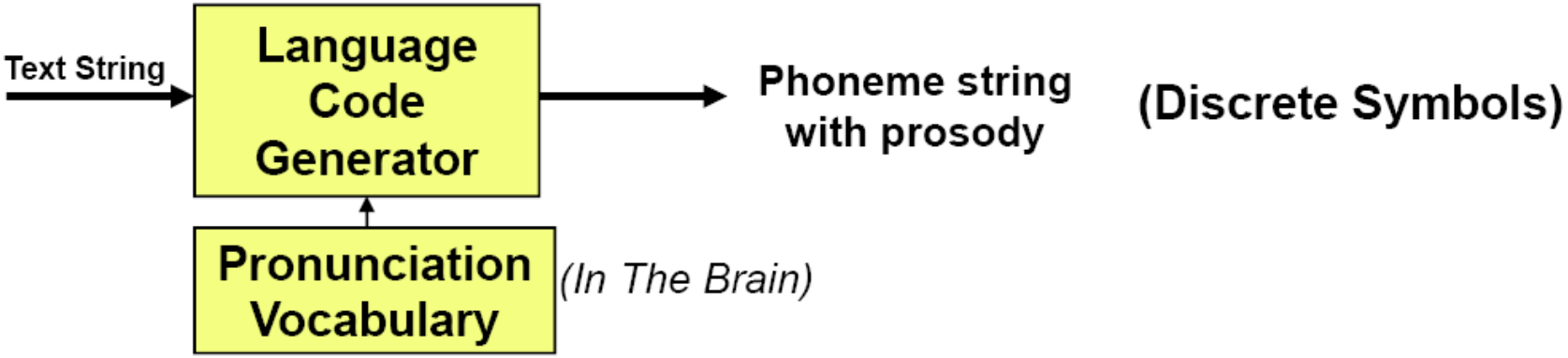
Speech Production Model

- Message Formulation 信息形成
  - desire to communicate an idea, a wish, a request, …
    ⟶ express the message as a sequence of words



Speech Production Model

- Language Code 语言编码
  - need to convert chosen text string to a sequence of sounds in the language that can be understood by others
  - need to give some form of emphasis, prosody (tune, melody) to the spoken sounds so as to impart non-speech information such as sense of urgency, importance, psychological state of talker, environmental factors (noise, echo)

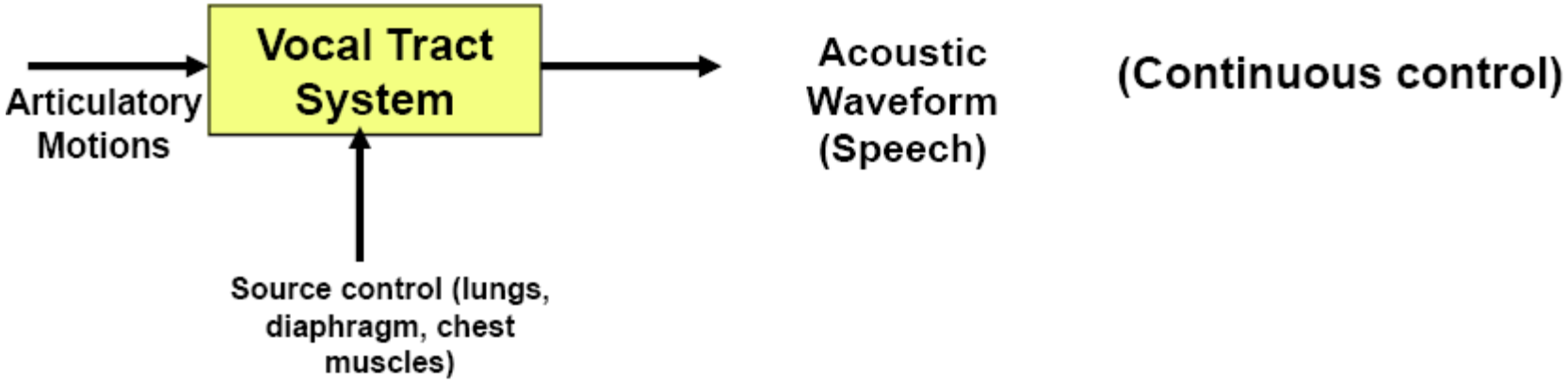


Speech Production Model

- Neuro-Muscular Controls 神经-肌肉控制
  - need to direct the neuro-muscular system to move the articulators (发音器官) (tongue, lips, teeth, jaws, velum(软腭)) so as to produce the desired spoken message in the desired manner

Speech Production Model

- Vocal Tract (声道) System
  - need to shape the human vocal tract system and provide the appropriate sound sources to create an acoustic waveform (speech) that is understandable in the environment in which it is spoken



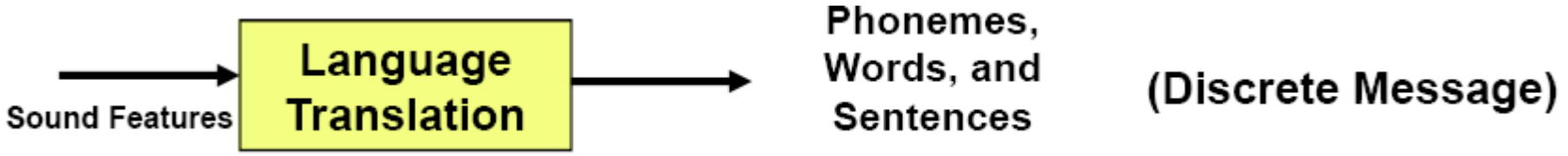
Speech Perception Model

- The acoustic waveform impinges(冲击) on the ear (the basilar membrane(基底膜)) and is spectrally analyzed by an equivalent filter bank(滤波器组) of the ear

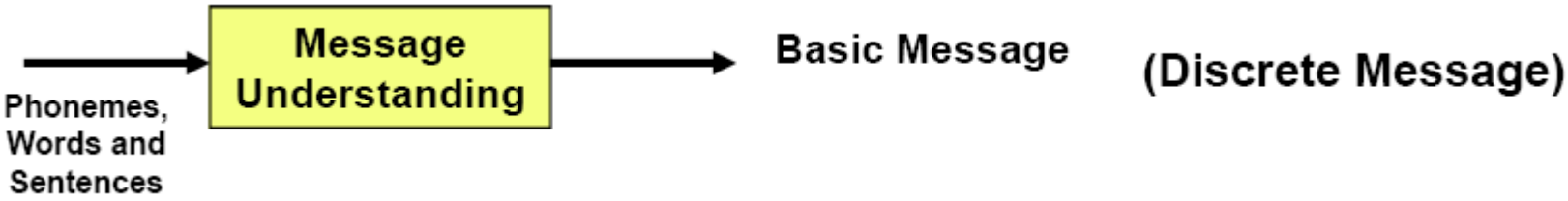
- The signal from the basilar membrane is neurally transduced and coded into features that can be decoded by the brain

Speech Perception Model

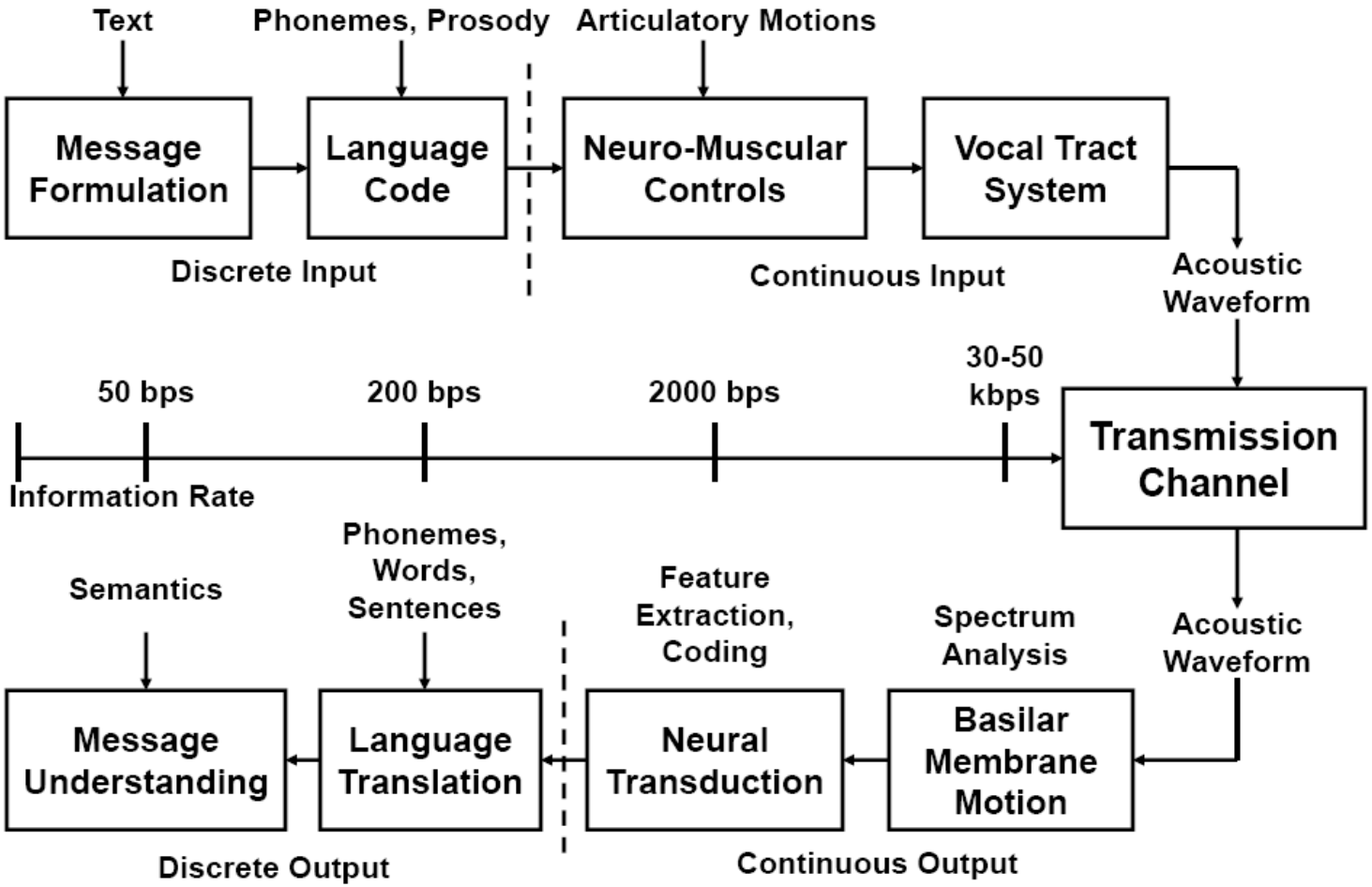
- The brain decodes the feature stream into sounds, words and sentences



- The brain determines the meaning of the words via a message understanding mechanism



The Speech Chain



Goal: Find out if your office mate has had lunch

Text: "Did you eat yet?"

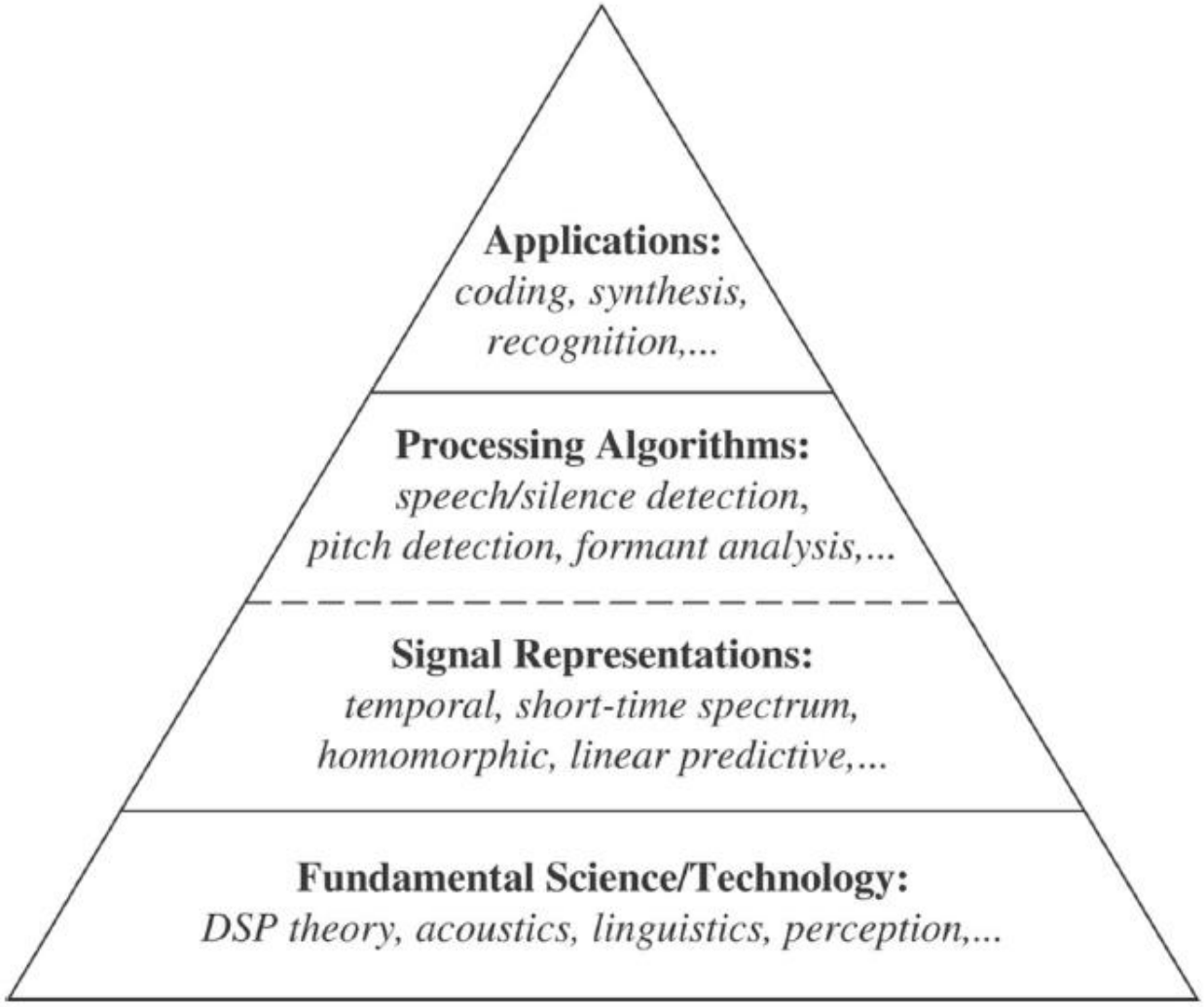
Phonemes: "did yu it yɛt?"

Articulator Dynamics: dI jə it jɛt

Information Rate of Speech

- Text (discrete)
  - 2^5 symbols, 10 symbols/s -> 50bps
- Phonemes & Prosody (discrete)
  - 200 bps
- Articulatory motions (continuous)
  - Relatively slow movement of articulators ~2000bps
- Acoustic waveform (continuous)
  - 64,000 bps ~ 705,600 bps

The Speech Stack



Speech Science(语音科学)

- **Linguistics**（语言学）: science of language, including syntax, semantics, phonetics, phonology, etc.
- **Syntax**（句法，语法）: analysis and description of the grammatical structure of a body of textual material
- **Semantics**（语义学）: analysis and description of the meaning of a body of textual material and its relationship to a task description of the language
- **Phonetics**（语音学）: study of speech sounds and their production, transmission, and perception, and their analysis, classification, and transcription
  - Articulatory/Acoustic/Auditory Phonetics
- **Phonology**（音系学）: systematic organization of sounds in languages, systems of phonemes in particular languages
- **Phonemes**（音位，音素）: smallest set of units considered to be the basic set of distinctive sounds of a languages (20-60 units for most languages)

Applications of Speech Signal Processing

Applications of Speech Signal Processing

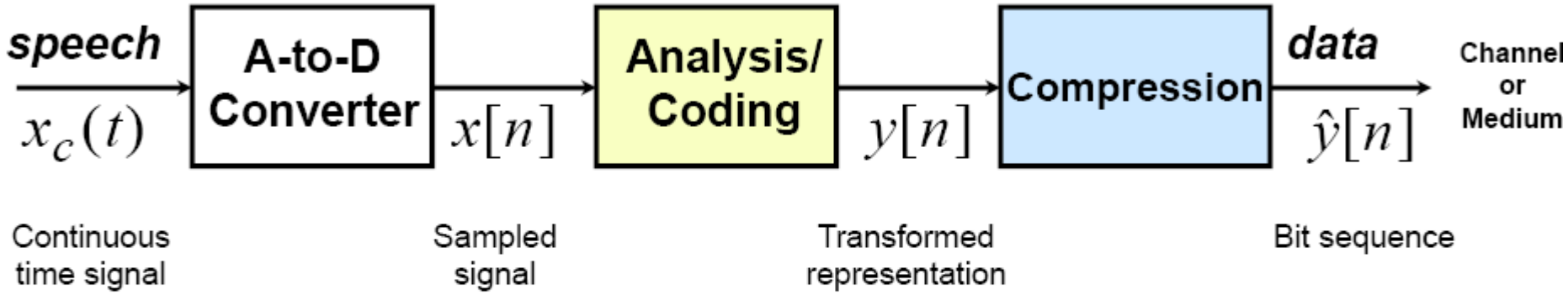
- Speech coding (语音编码)
- Speech synthesis (语音合成)
- Speech recognition and understanding (语音识别与理解)
- Other speech applications

Speech Coding

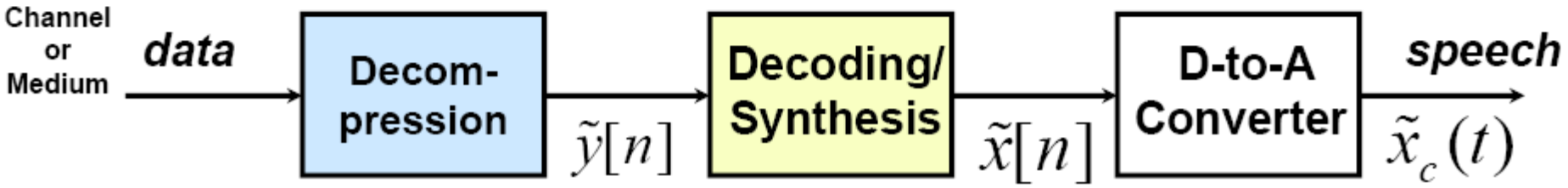
- The process of transforming a speech signal into a representation for efficient transmission and storage of speech
  - narrowband and broadband wired telephony
  - cellular communications
  - Voice over IP (VoIP) to utilize the Internet as a real-time communications medium
  - secure voice for privacy and encryption for national security applications
  - extremely narrowband communications channels, e.g., battlefield applications using HF radio
  - storage of speech for telephone answering machines, IVR systems, prerecorded messages

Speech Coding

Encoding



Decoding



Speech Synthesis

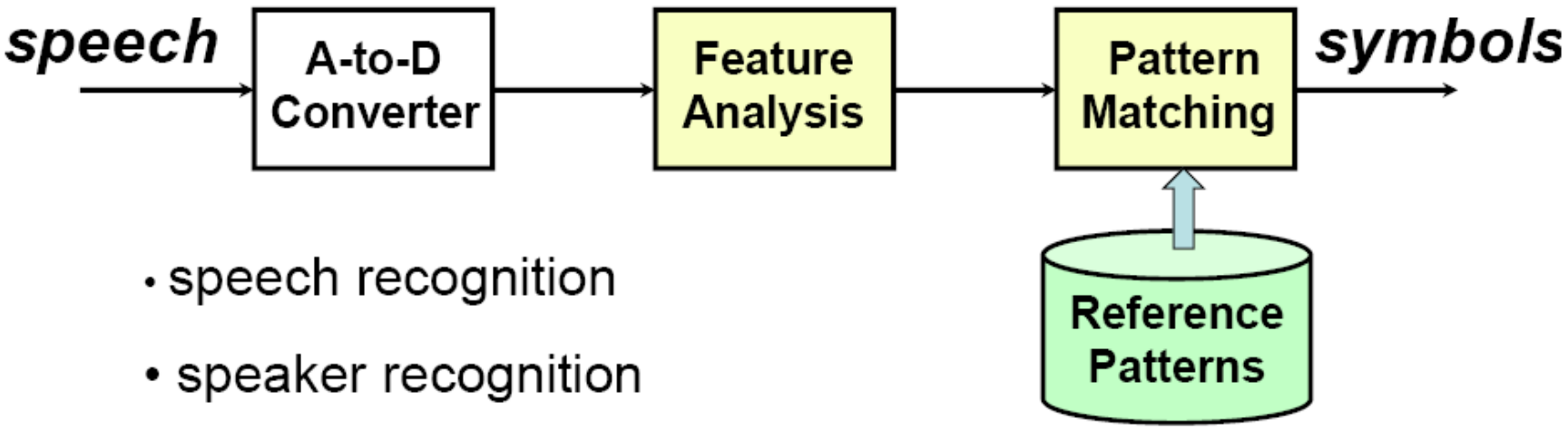
- The process of generating a speech signal using computational means for effective human-machine interactions
  - machine reading of text or email messages
  - telematics feedback in automobiles
  - talking agents for automatic transactions
  - automatic agent in customer care call center
  - handheld devices such as foreign language phrasebooks, dictionaries, crossword puzzle helpers
  - announcement machines that provide information such as stock quotes, airlines
  - schedules, weather reports, etc.

Speech Synthesis

Speech Recognition and Understanding

- The process of extracting usable linguistic information from a speech signal in support of human-machine communication by voice
  - command and control (C&C) applications, e.g., simple commands for spreadsheets, presentation graphics, appliances
  - voice dictation to create letters, memos, and other documents
  - natural language voice dialogues with machines to enable Help desks, Call Centers
  - voice dialing for cellphones and from PDA's and other small devices
  - agent services such as calendar entry and update, address list modification and entry, etc.

Pattern Matching Problems



- speech recognition
- speaker recognition
- speaker verification
- word spotting
- automatic indexing of speech recordings

Other Speech Applications

- Speaker Verification (话者确认)
  - for secure access to premises, information, virtual spaces
- Speaker Recognition (话者识别)
  - for legal and forensic purposes—national security; also for personalized services
- Speech Enhancement (语音增强)
  - for use in noisy environments, to eliminate echo, to align voices with video segments, to change voice qualities, to speed-up or slow-down prerecorded speech (e.g., talking books, rapid review of material, careful scrutinizing of spoken material, etc)
  - potentially to improve intelligibility and naturalness of speech
- Language Translation (语言翻译)
  - to convert spoken words in one language to another to facilitate natural language dialogues between people speaking different languages, i.e., tourists, business people

History of Speech Signal Processing

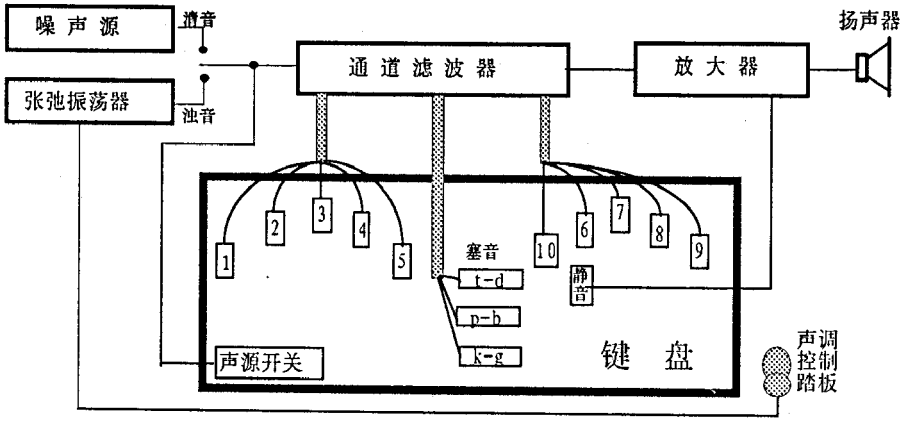
History of Speech Signal Processing

- Invention of telephone, Bell 1876
  - “Watson, if I can get a mechanism which will make a current of electricity vary its intensity as the air varies in density when sound is passing through it, I can telegraph any sound, even the sound of speech”

History of Speech Signal Processing

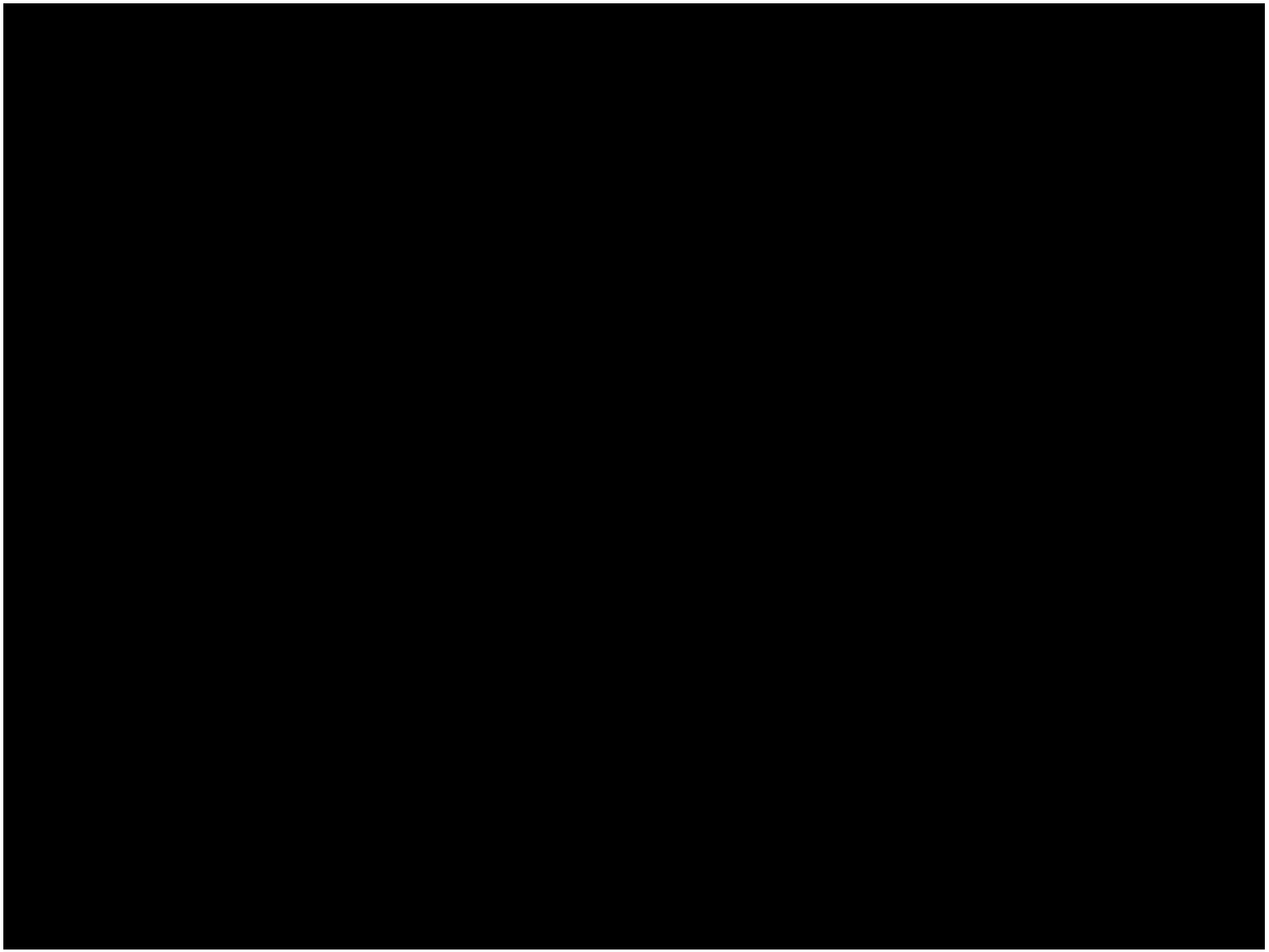
- VOCODER and VODER, Dudley
  - VOCODER (VOice enCODER) 声码器
    - a method of reproducing speech through electronic means
    - source-filter model
    - use parallel band-pass filter to filter speech into ten specific audio spectrum bands, rendering it more easily transmitted over telephone lines
  - VODER (Voice Operation DEmonstratoR)
    - a console from which an operator could create phrases of speech controlling a VOCODER with a keyboard and foot pedals (踏板)
    - 1939 World Fair in NYC

VODER

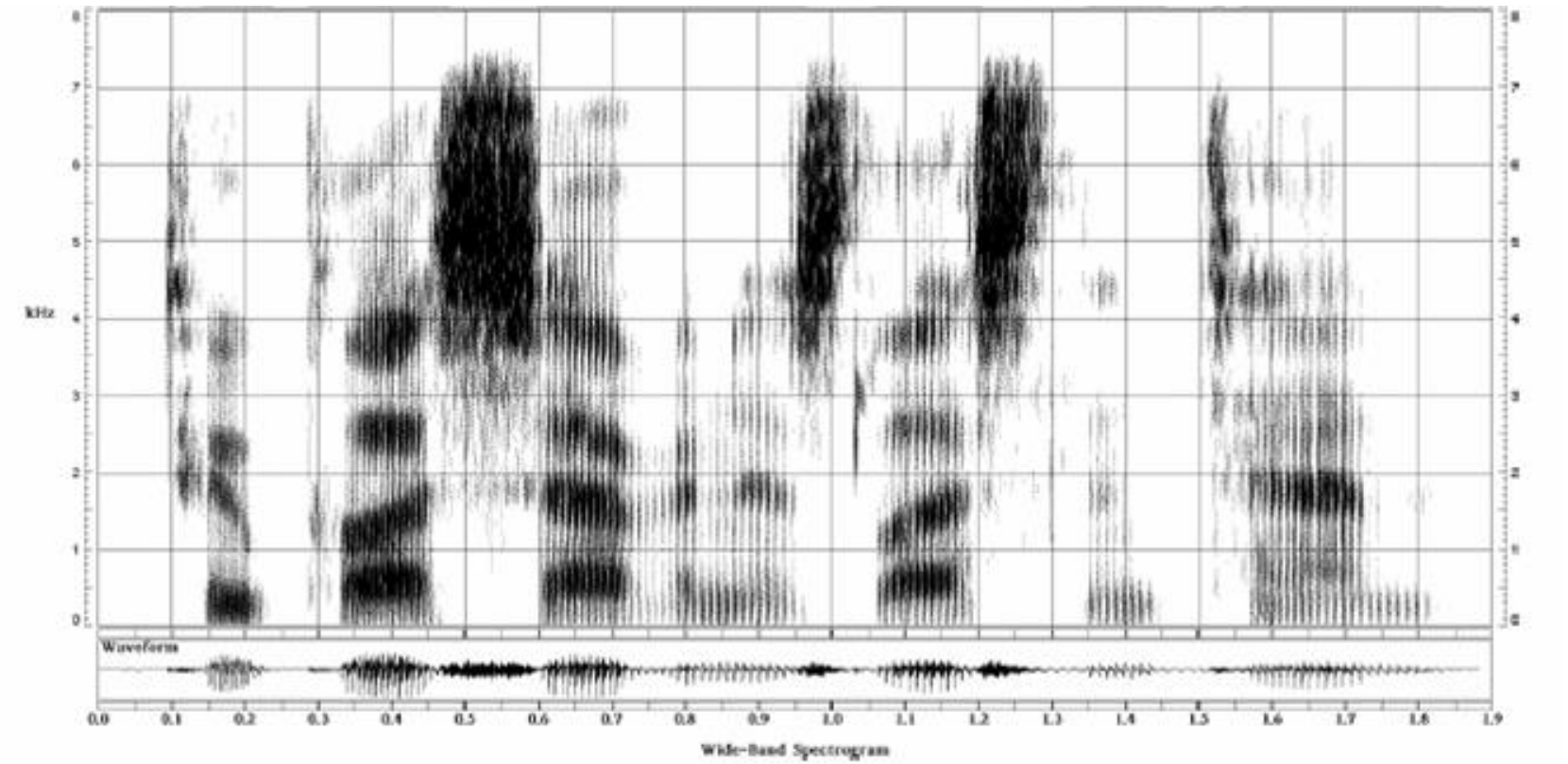


图一 "Voder" 简图(引自 Dudley H. 1939)

VODER

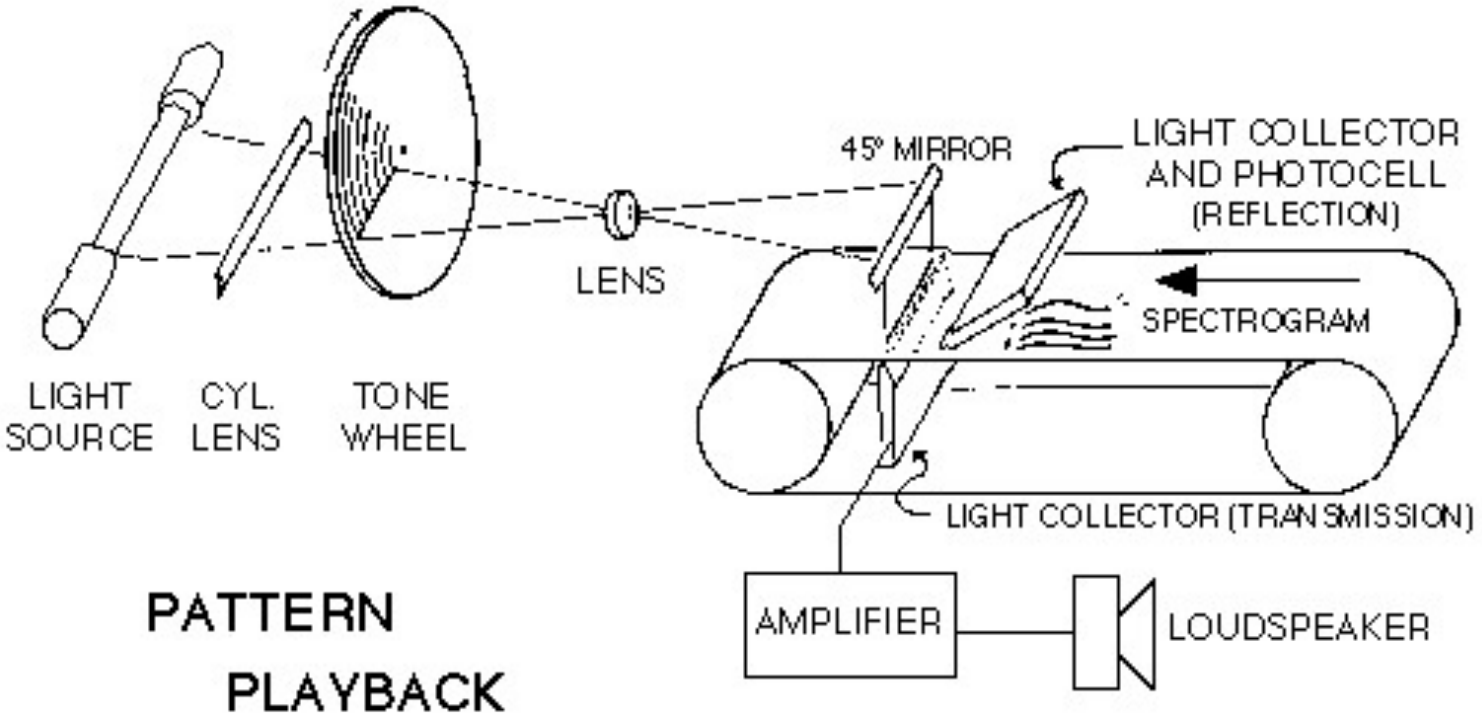


Sound Spectrograph (语谱仪), Bell Lab, 1947

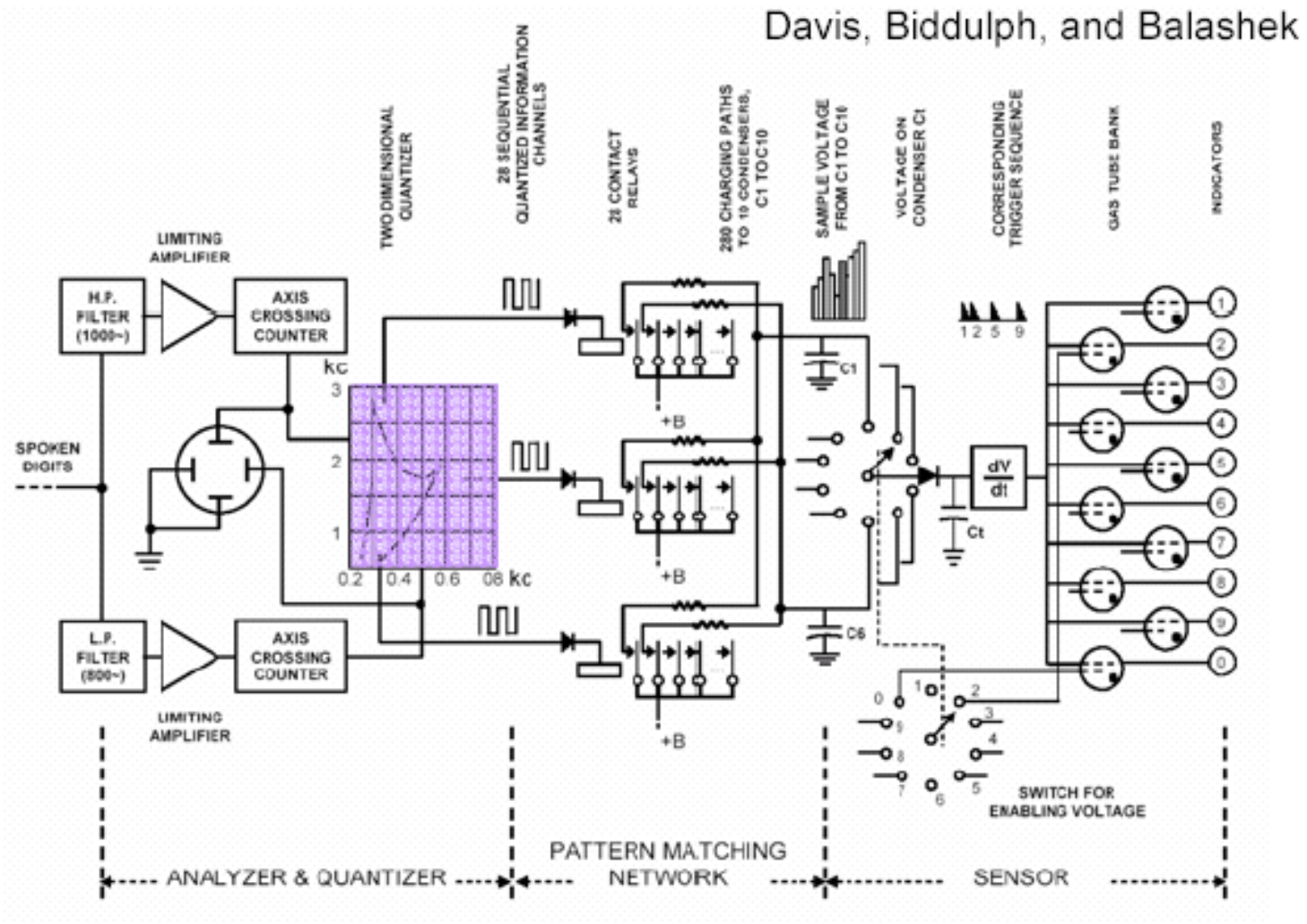


Two plus seven is less than ten

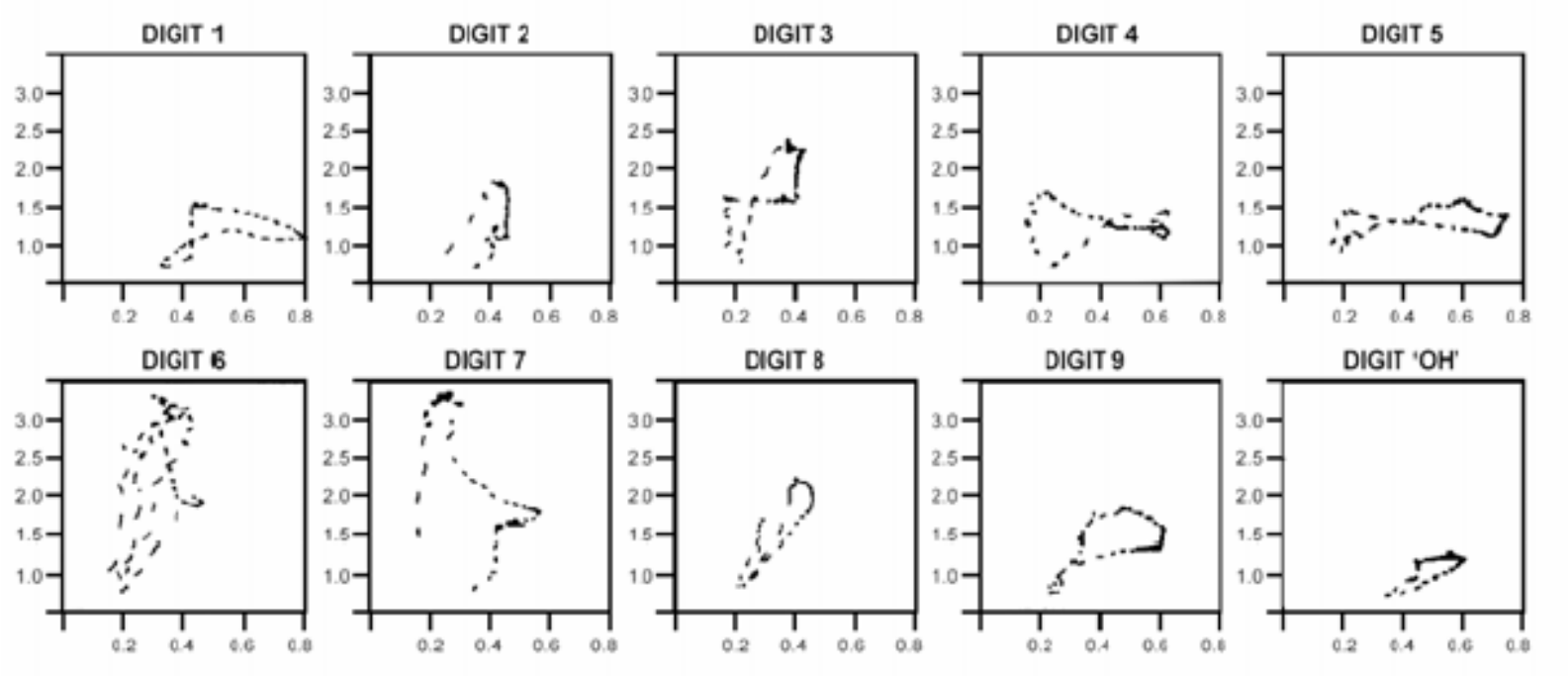
Pattern Playback, Haskins Lab, 1950



Digit Recognizer, Bell Labs, 1952



Digit Pattern



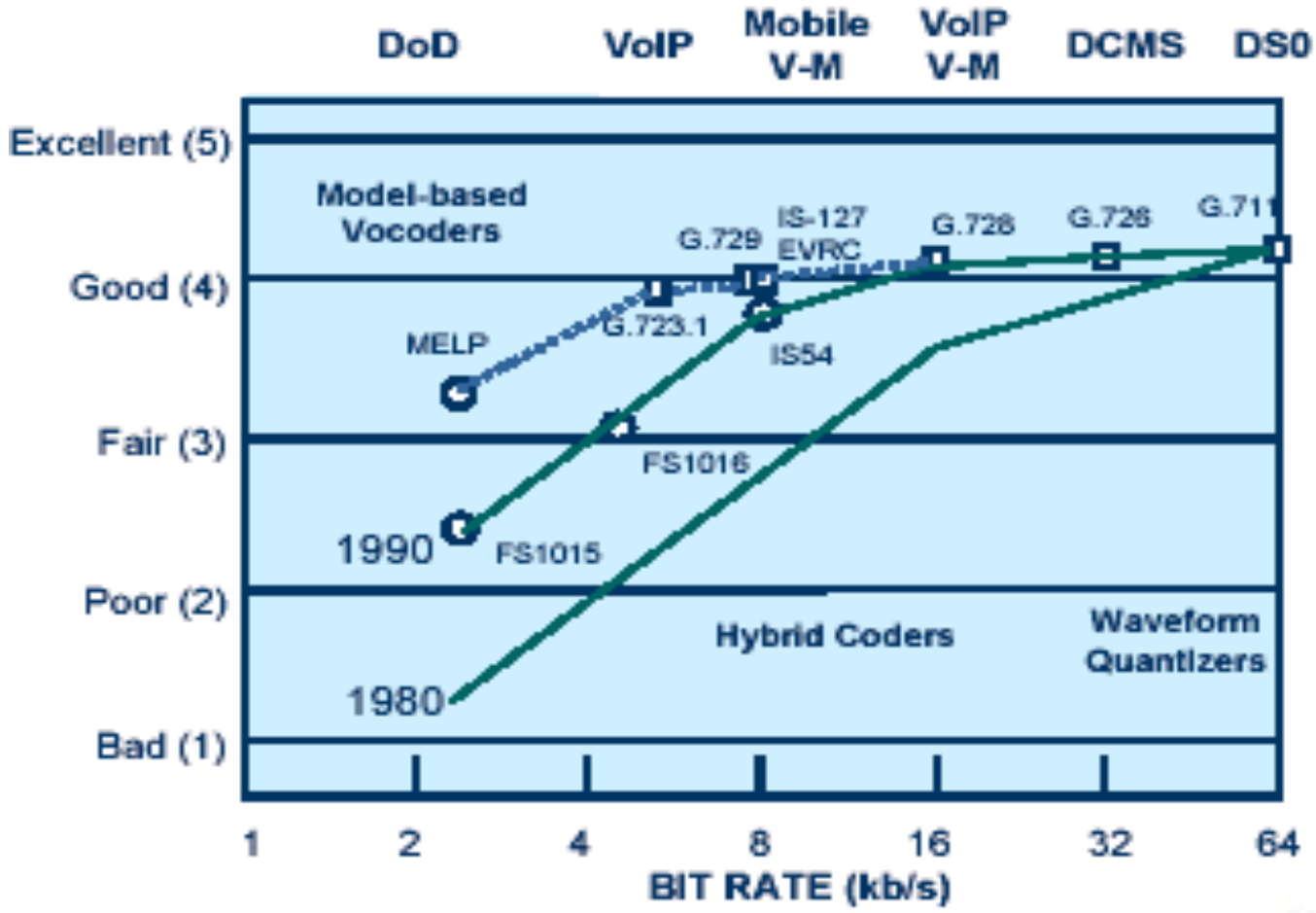
The idea was to track the first two formants.

1960-70’s

- Fant, “Acoustic Theory of Speech Production”, 1970
- Breakthrough in DSP since the mid 1960’
  - 1965 FFT
  - 1968 Homomorphic Processing (同态处理)
  - mid 1970’s Linear Prediction Analysis (线性预测分析)
  - late 1970’s Vector Quantization (矢量量化)
- Pattern matching techniques
  - 1970’s Dynamic Time Warping (动态时间规整)
- Widely application of computers
- DARPA started Speech Understanding Research (SUR) program in 1970’s

Since 1980's

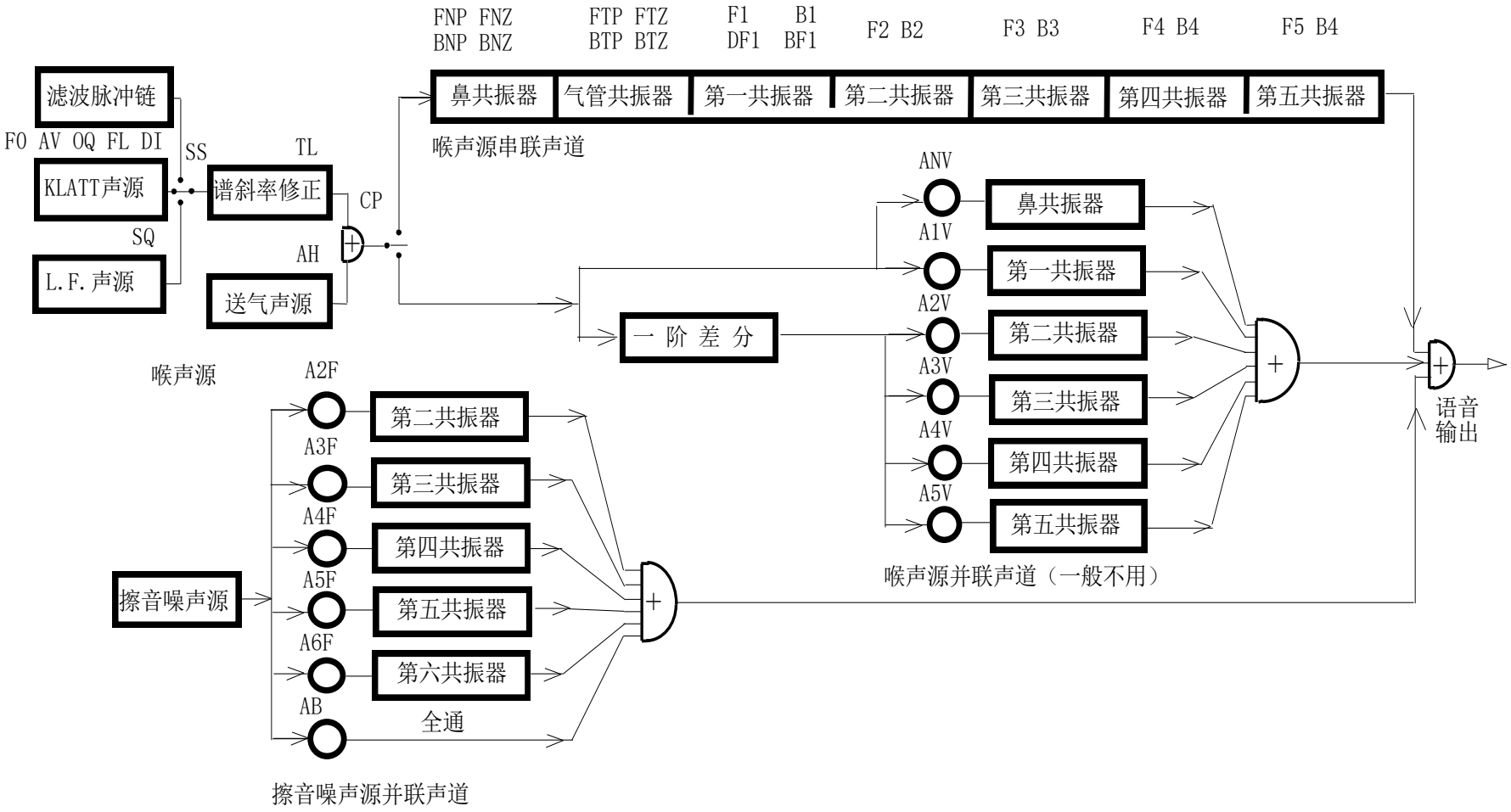
- Speech Coding
  - 1980 LPC-10 2.4kbps
  - 1988 FS-1016 4.8kbps
  - 1990's MBE 2.4kbps
  - ITU-T G-series standard, model-based VOCODER


DoD
VoIP
Mobile V-M
VoIP V-M
DCMS
DS0
Excellent (5)
Good (4)
Fair (3)
Poor (2)
Bad (1)
Model-based Vocoders
MELP
G.723.1
G.729
IS-127 EVRC
IS54
G.728
G.726
G.711
FS1016
1990
FS1015
1980
Hybrid Coders
Waveform Quantizers
1
2
4
8
16
32
64
BIT RATE (kb/s)

Since 1980's

- Speech synthesis
  - 1980 Klatt cascade/parallel formant synthesizer
  - Waveform concatenation
    - rule-based, TD-PSOLA
    - corpus-based , unit selection
  - HMM-based parametric speech synthesis

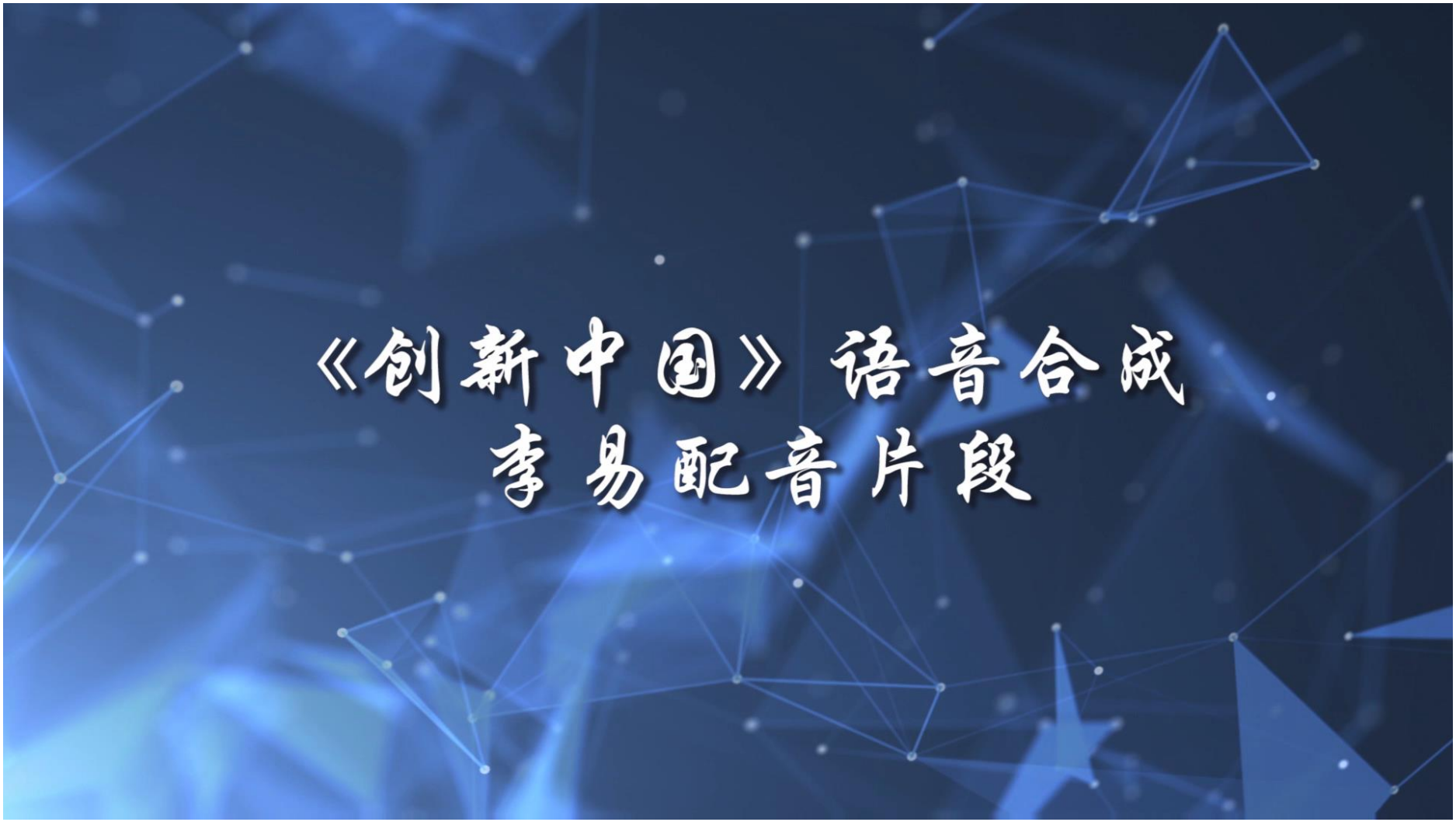
Klatt Synthesizer



Waveform Concatenation Synthesis - iFLYTEK

年份	1995年	1998年	1999年	2001年	2003年
自然度	<3.0	3.0	3.5	3.8	4.3





《创新中国》语音合成
李易配音片段

Since 1980's

- Speech recognition
  - HMM-based Statistical pattern recognition framework
  - Development of VLSI and computer technology
  - Speech recognition systems
    - 1985 IBM "Tangora", isolated-word speech recognizer
    - 1990 IBM "Dragon Dictate", first large-vocabulary speech-to-text system for general-purpose dictation
    - 1990's CMU "Sphinx", continuous-speech, speaker-independent recognition system
    - 1997 IBM "ViaVoice"

IBM

1997年9月
发布Viavoice语音识别软件中文版，从上个世纪70年代开始进行语音技术研究

Microsoft

2007年3月
以8亿美金价格收购语音搜索业务公司TellMe，加大对语音技术投入

2009年10月
微软发布WIN7操作系统，集成语音识别技术

Tellme.
A Microsoft Subsidiary

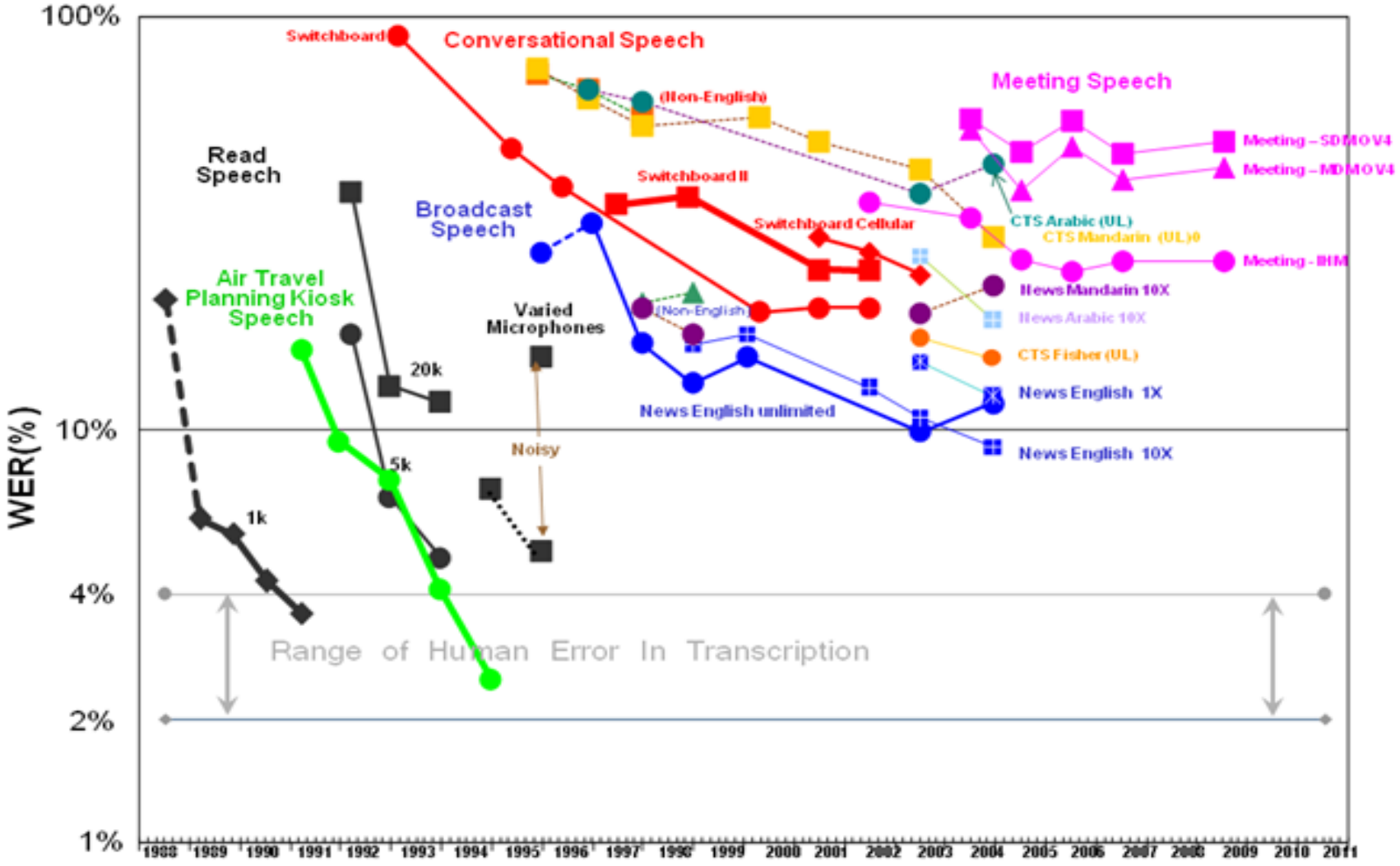
Google

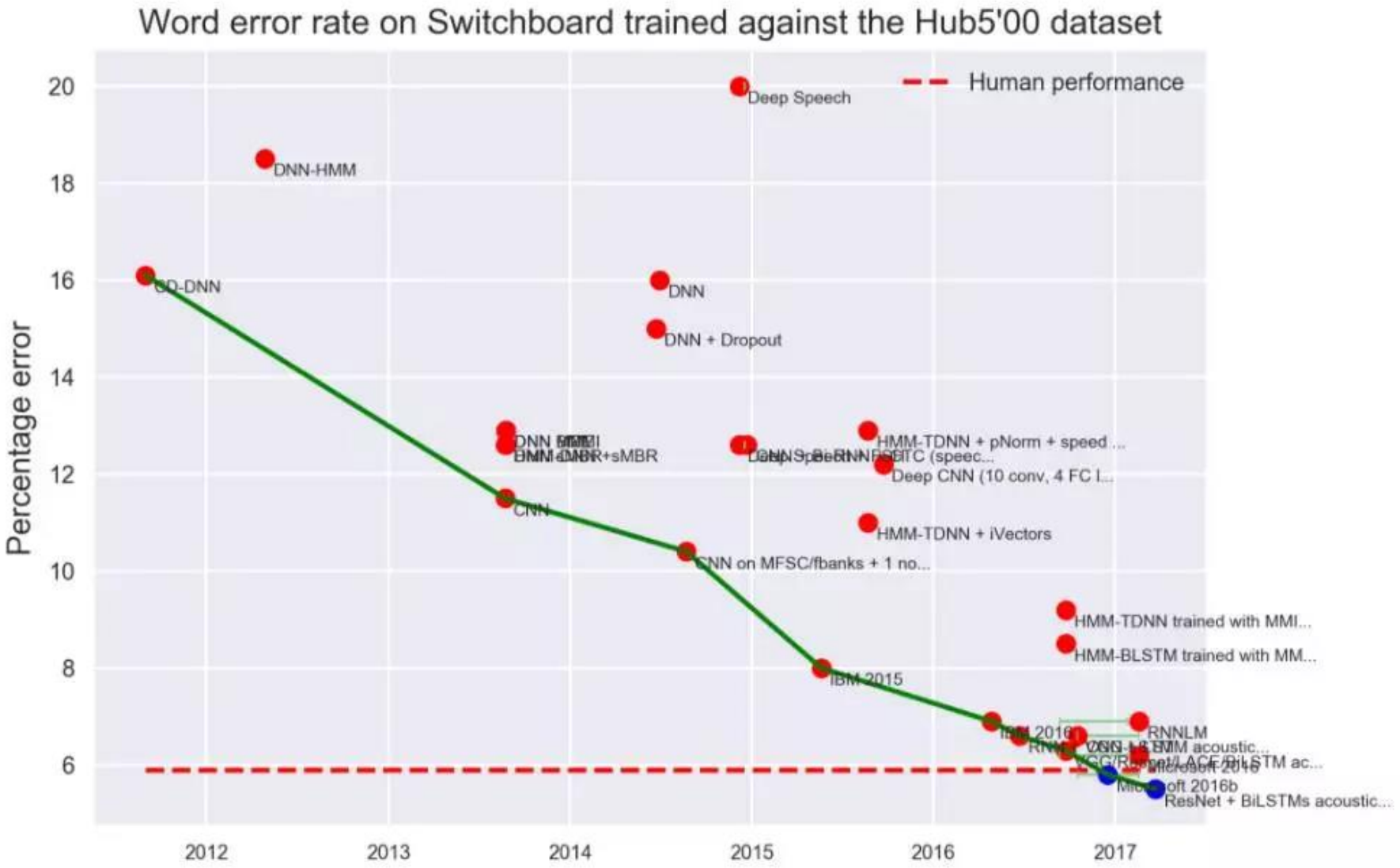
2007-2010年
先后发布电话语音搜索，互联网移动语音搜索，Google Voice Action

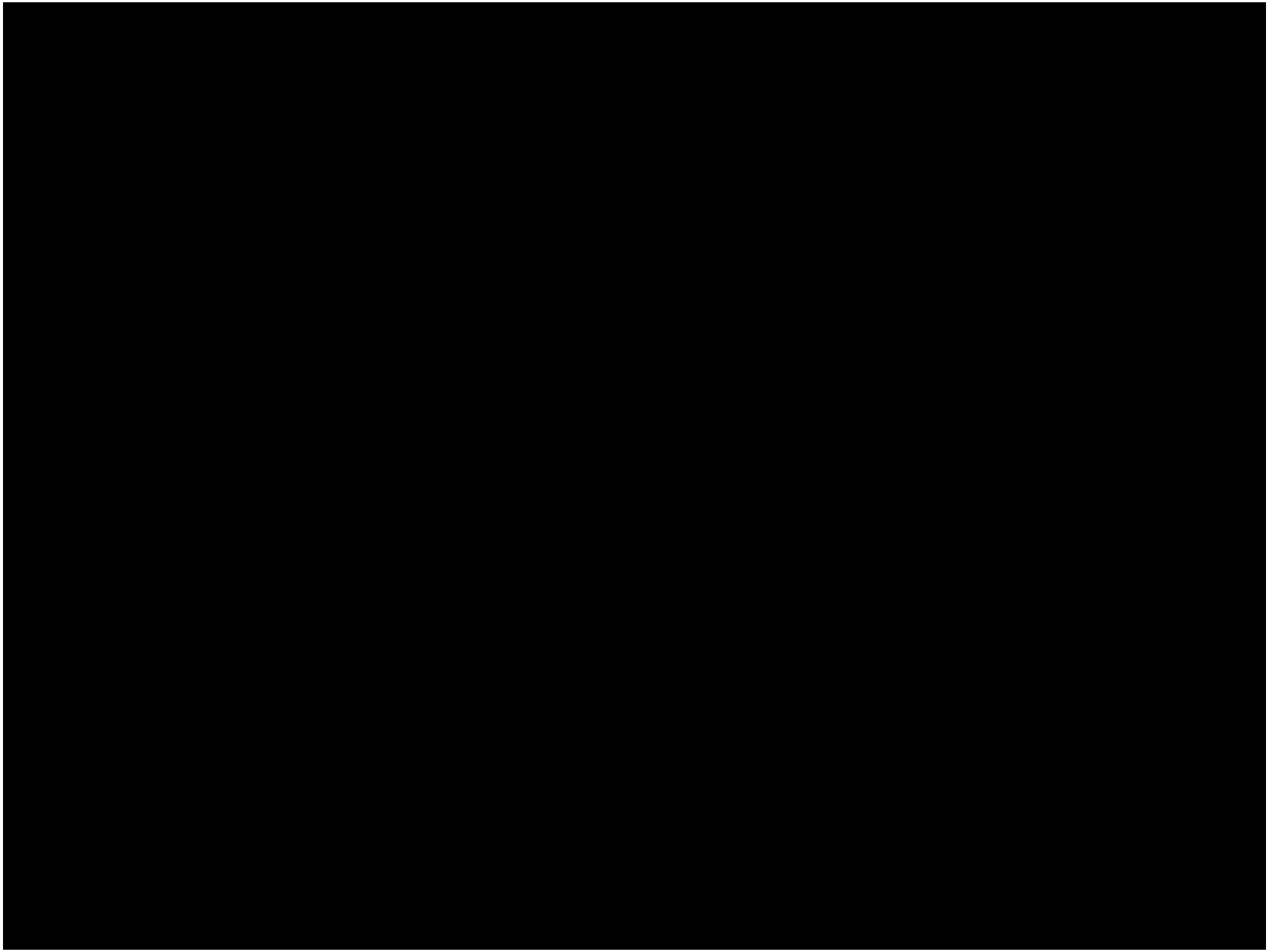
2010年4月
收购语音服务提供商Siri，宣布将在iPhone中提供智能语音服务

The History of Automatic Speech Recognition Evaluations at NIST

NIST STT Benchmark Test History – May. '09




Word error rate on Switchboard trained against the Hub5'00 dataset
Percentage error
20
18
16
14
12
10
8
6
2012
2013
2014
2015
2016
2017
Human performance
Deep Speech
DNN-HMM
CD-DNN
DNN
DNN + Dropout
HMM-TDNN + pNorm + speed ...
Deep CNN (10 conv, 4 FC l...
HMM-TDNN + iVectors
CNN
CNN on MFSC/fbanks + 1 no...
HMM-TDNN trained with MMI...
HMM-BLSTM trained with MM...
IBM 2015
RNNLM
ResNet + BiLSTMs acoustic...



What We Will Be Learning

- review some basic DSP concepts
- speech production model—acoustics, articulatory concepts, speech production models
- speech perception model—ear models, auditory signal processing
- time domain processing concepts—speech properties, pitch, voiced-unvoiced, energy, autocorrelation, zero-crossing rates
- short time Fourier analysis methods—digital filter banks, spectrograms, analysis-synthesis systems, vocoders
- homomorphic speech processing—cepstrum, pitch detection, formant estimation, homomorphic vocoder
- linear predictive coding methods—autocorrelation method, covariance method, lattice methods, relation to vocal tract models
- speech waveform coding and source models—delta modulation, PCM, mu-law, ADPCM, vector quantization, multipulse coding, CELP coding
- methods for speech synthesis and text-to-speech systems—physical models, formant models, articulatory models, concatenative models
- methods for speech recognition—the Hidden Markov Model (HMM)