同济大学

Introduction to ASR

SHEN Ying
SSE, Tongji University

Introduction to this course

INSTRUCTOR

SHEN Ying (yingshen@tongji.edu.cn)

Room 509R, Jishi Building

TA

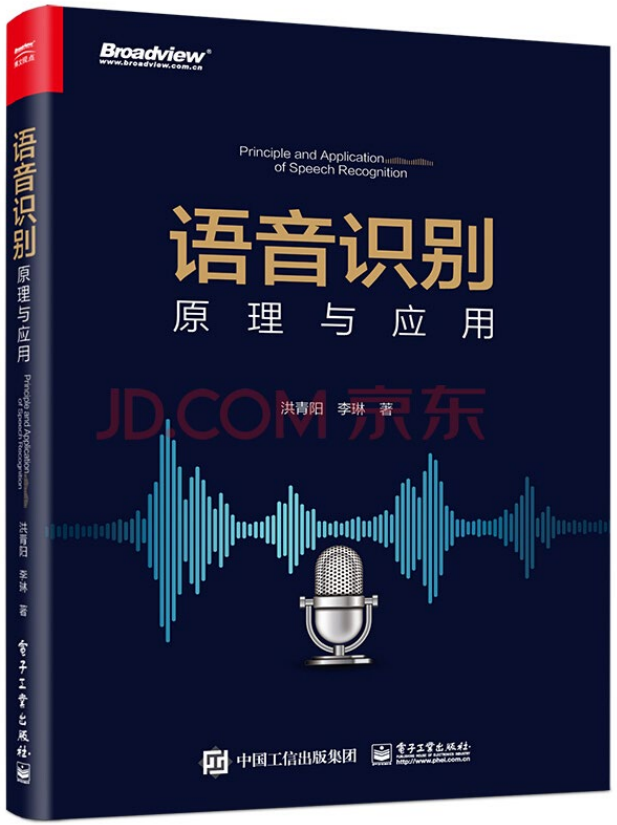
LI Yusen

Room 308L, Jishi Building

Introduction to this course

TEXTBOOK

语音识别 原理与应用



Daniel Jurafsky and James H. Martin (2008). *Speech and Language Processing*, (3nd edition draft). https://web.stanford.edu/~jurafsky/slp3/

26: Automatic Speech Recognition and Text-to-Speech

Appendix Chapters (will be just on the web)

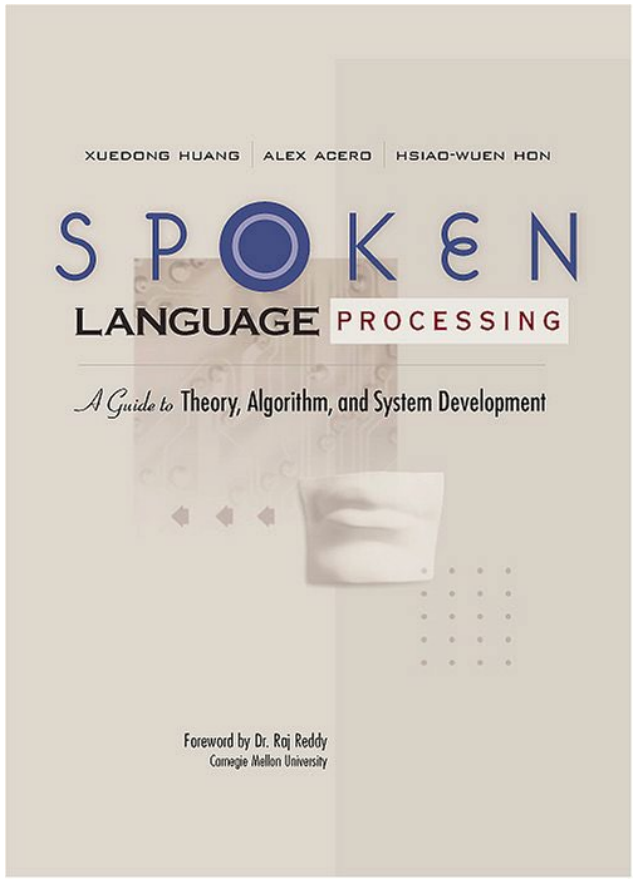
A: Hidden Markov Models

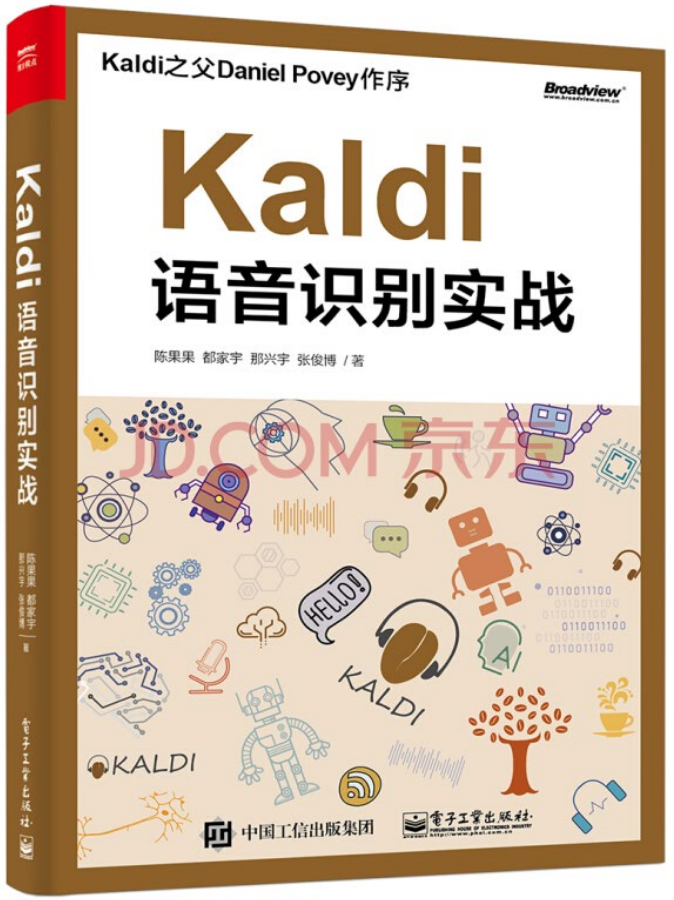
B: Spelling Correction and the Noisy Channel

C: Statistical Constituency Parsing

Introduction to this course

OTHER MATERIALS





Introduction to this course

REVIEW AND TUTORIAL ARTICLES

G&Y: MJF Gales and SJ Young (2007). The Application of Hidden Markov Models in Speech Recognition, *Foundations and Trends in Signal Processing*, **1** (3), 195-304.

S Young (1996). A review of large-vocabulary continuous-speech recognition, IEEE Signal Processing Magazine 13 (5), 45-57.

R&H:S Renals and T Hain (2010). Speech Recognition, in *Computational Linguistics and Natural Language Processing Handbook*, A Clark, C Fox and S Lappin (eds.), Blackwells, chapter 12, 299-332.

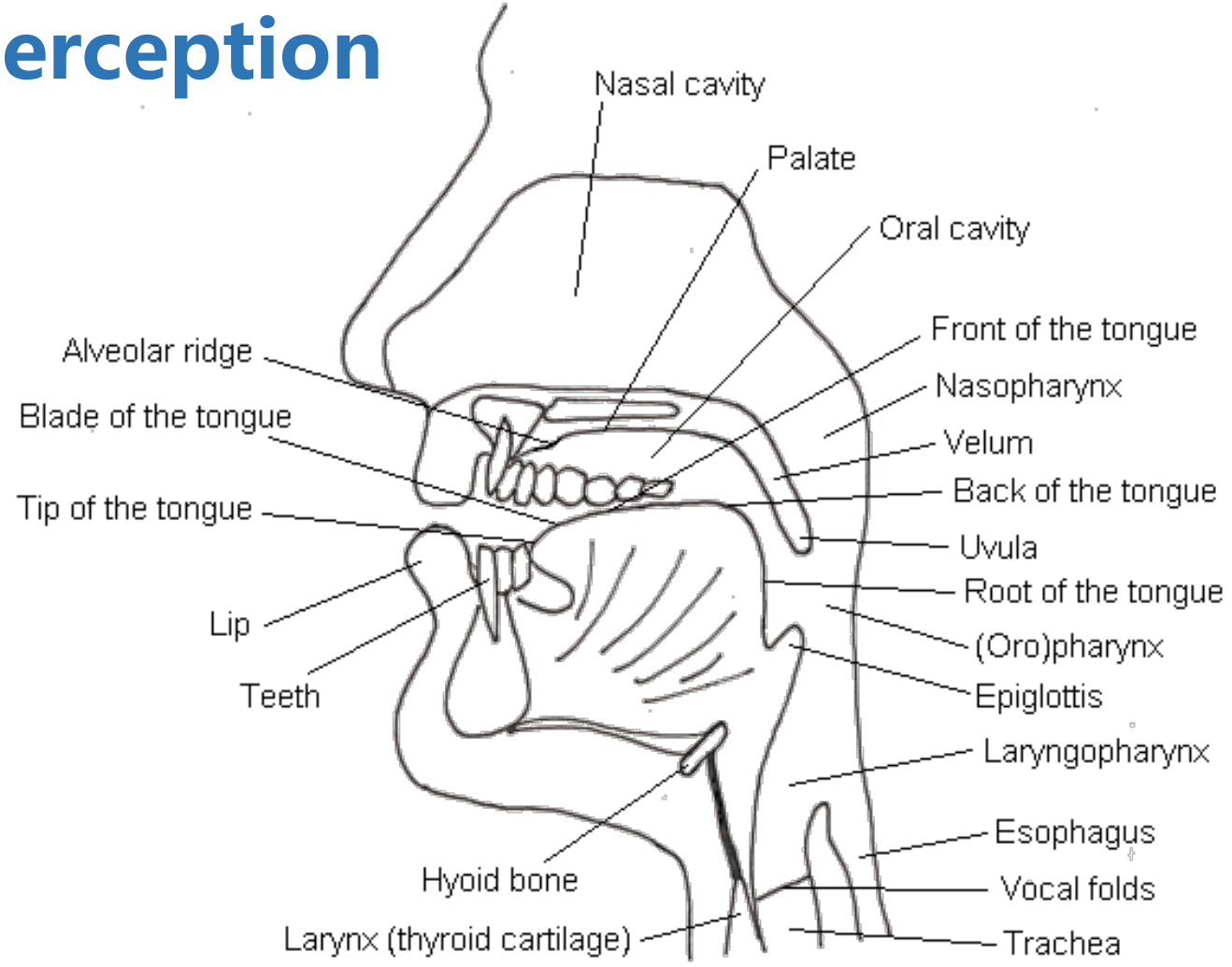
G Hinton et al (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, IEEE Signal Processing Magazine, **29**(6):82-97.

S Young (2008). HMMs and Related Speech Recognition Technologies, in *Springer Handbook of Speech Processing*, J Benesty, MM Sondhi and Y Huang (eds), chapter 27, 539-557.

目 录

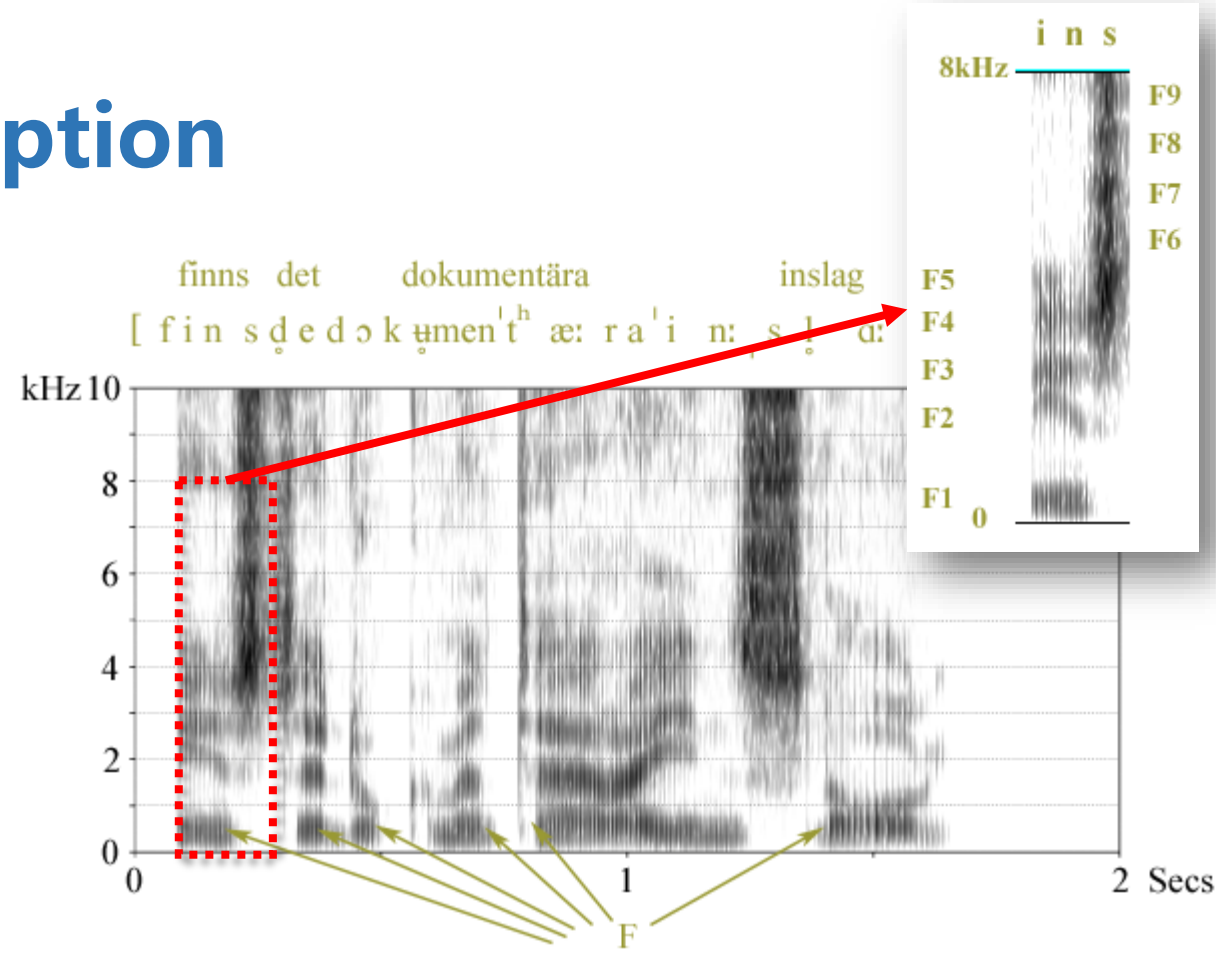
Voice production and perception

- Human vocal organs include: lungs, trachea, vocal folds, larynx, pharynx, nasal cavity, oral cavity and lips
- A person's voice pitch is determined by the resonant frequency of the vocal folds. In an adult male this frequency averages about 125 Hz, adult females around 210, in children the frequency is over 300 Hz

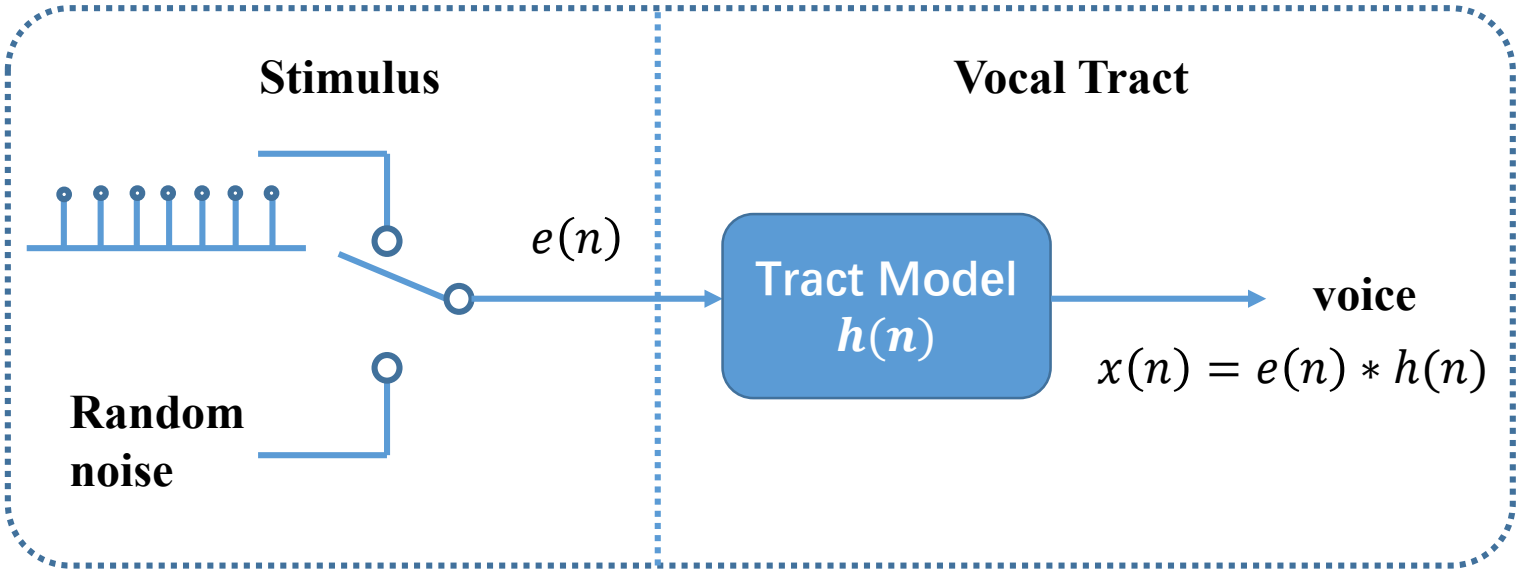


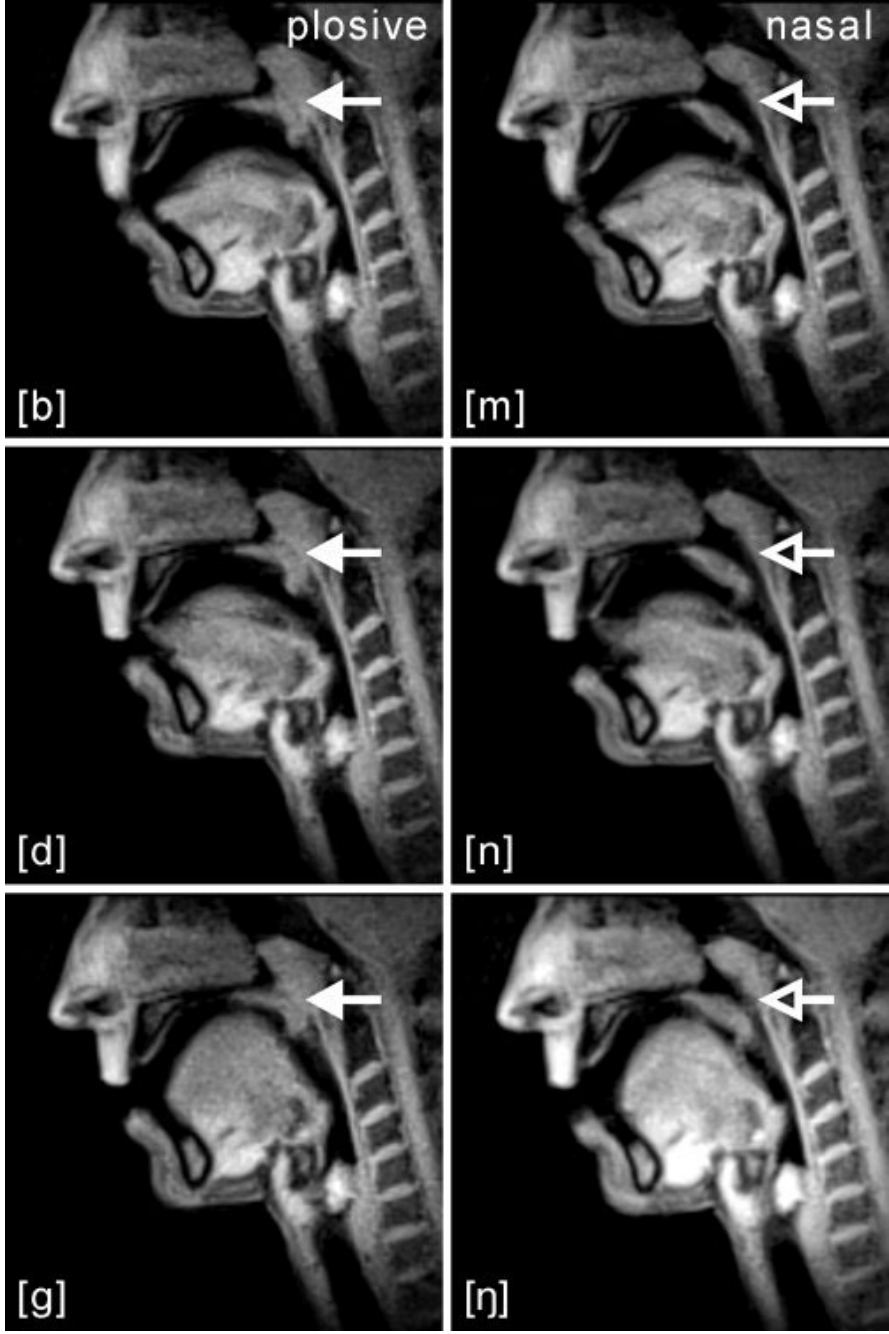
Voice production and perception

- The pharynx, oral cavity and nasal cavity are collectively referred to as the vocal tract
- A formant is a concentration of acoustic energy around a particular frequency in the speech wave
- formants occur at roughly 1000Hz intervals
- Each formant corresponds to a resonance in the vocal tract



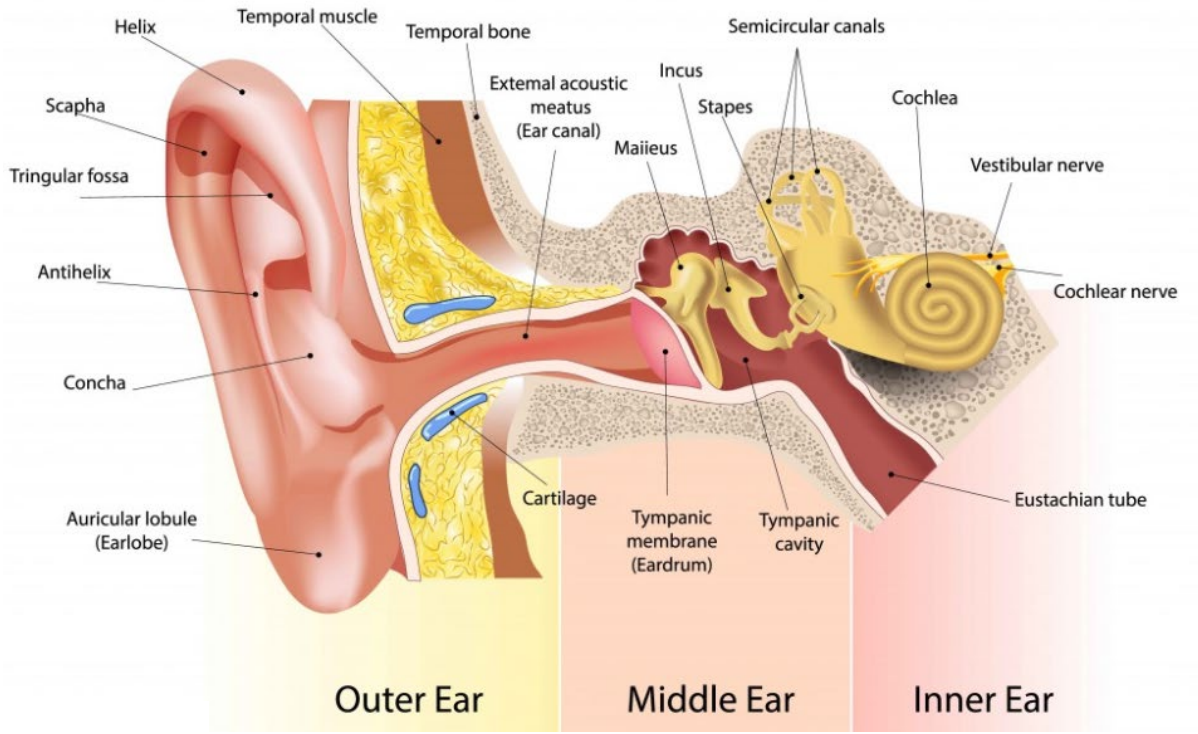
Voice production and perception

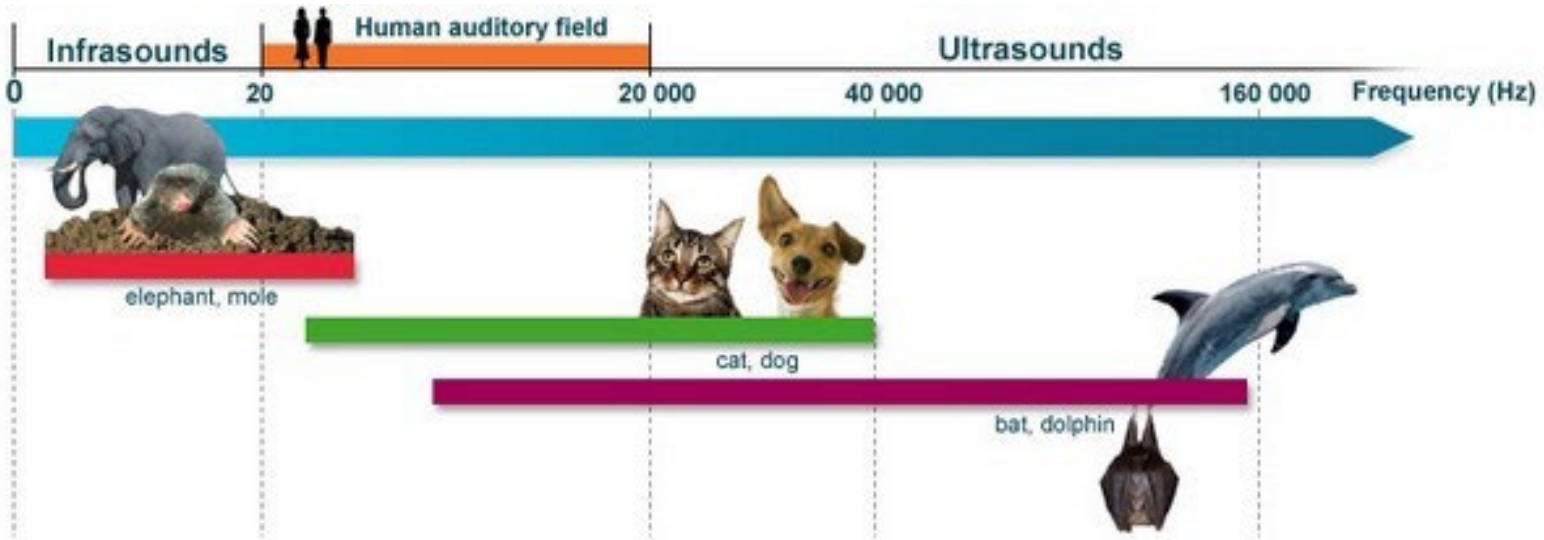




Voice production and perception

- The ear is the organ of hearing. It enables us to perceive and distinguish sounds.
- Human ear perceives frequencies between 20 Hz (lowest pitch) to 20 kHz (highest pitch)





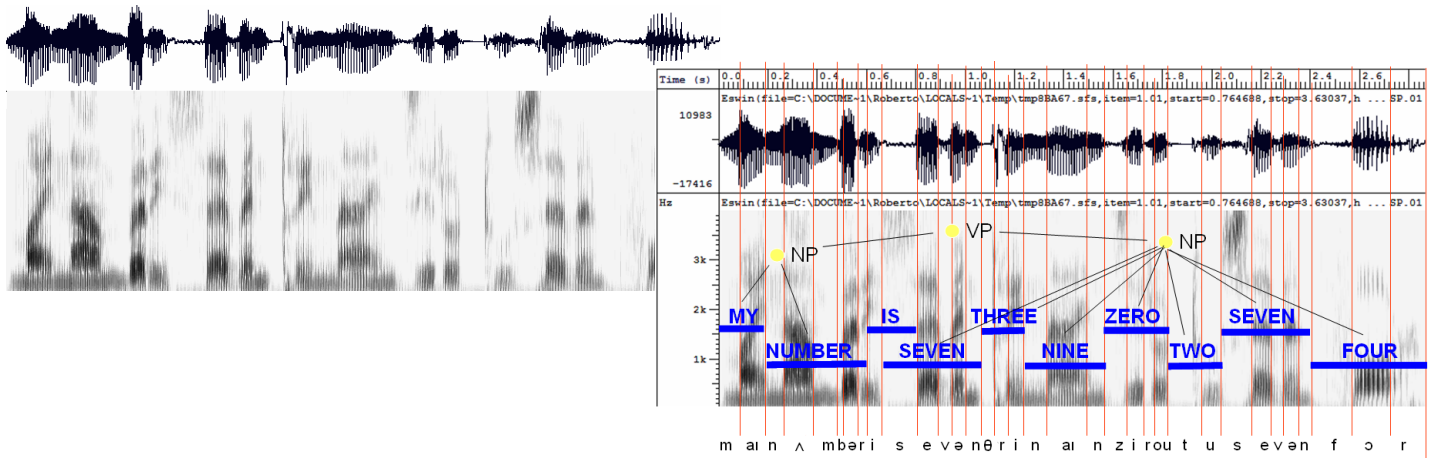
目 录

How might computers do it?

- Digitization
- Acoustic analysis of the speech signal
- Linguistic interpretation

Acoustic waveform

Acoustic signal



Speech recognition

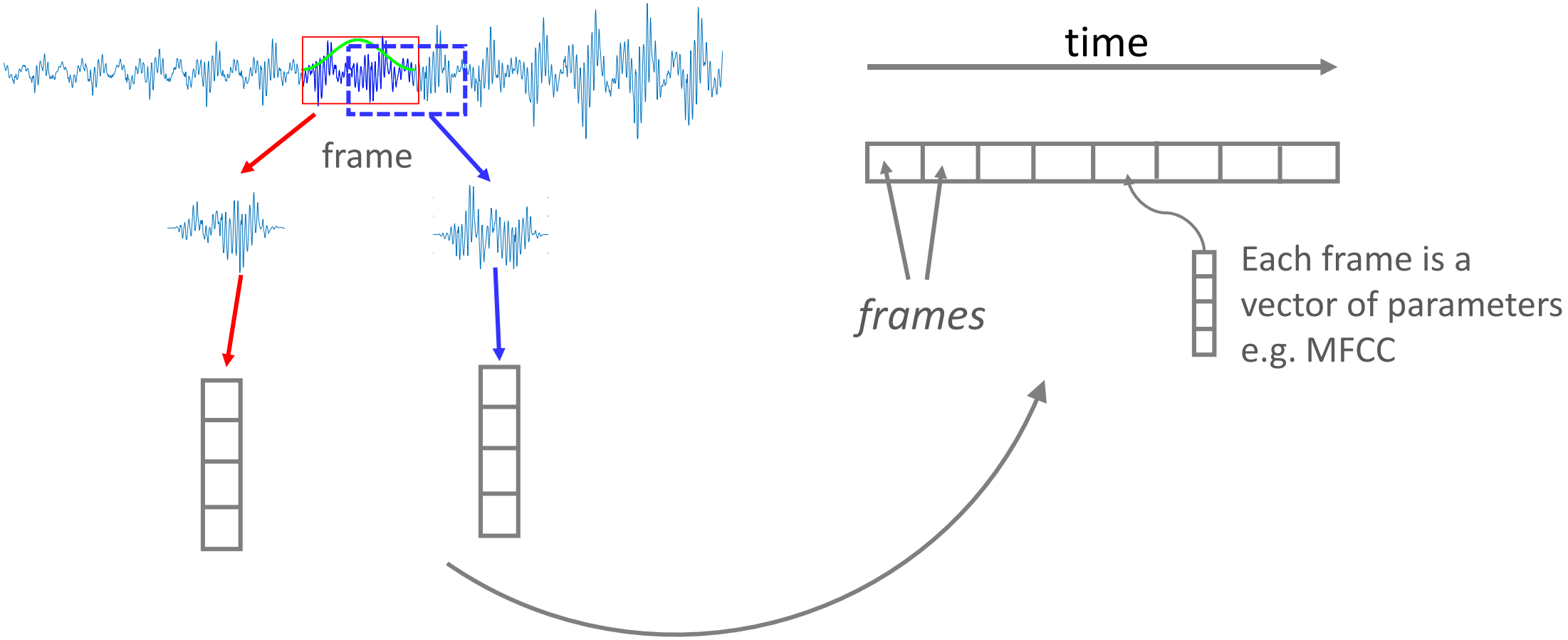
What is speech recognition?

- Speech-to-text transcription
  - Transform recorded audio into a sequence of words
  - Just the words, no meaning.... But do need to deal with acoustic ambiguity: “Recognize speech?” or “Wreck a nice beach?"
  - Speaker diarization: Who spoke when?
  - Speech recognition: what did they say?
  - Paralinguistic aspects: how did they say it? (timing, intonation, voice quality)
  - Speech understanding: what does it mean?

The speech recognition problem

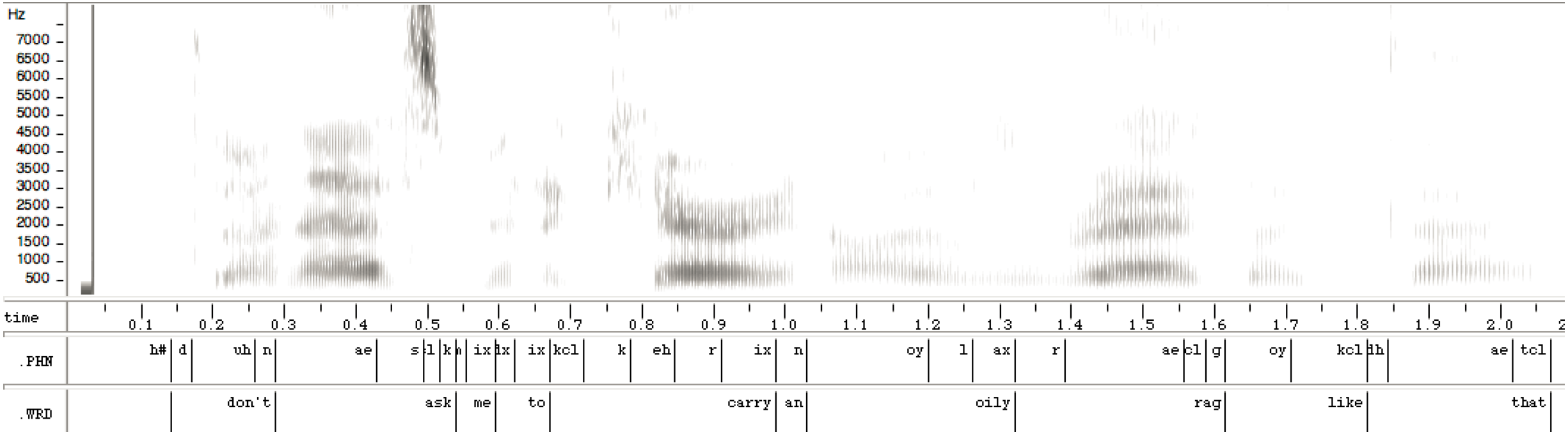
- We generally represent recorded speech as a sequence of acoustic feature vectors (observations), X and the output word sequence as W
- At recognition time, our aim is to find the most likely W, given X
- To achieve this, statistical models are trained using a corpus of labelled training utterances (X^n, W^n)

Representing recorded speech (X)



Represent a recorded utterance as a sequence of feature vectors

Labelling speech (W)



Labels may be at different levels: words, phones, etc.

Labels may be time-aligned – i.e. the start and end time of an acoustic segment corresponding to a label are known

Reading: Jurafsky & Martin chapter 7 (especially sections 7.4, 7.5)

History of ASR

第一阶段

模板匹配

第二阶段

统计模型

第三阶段

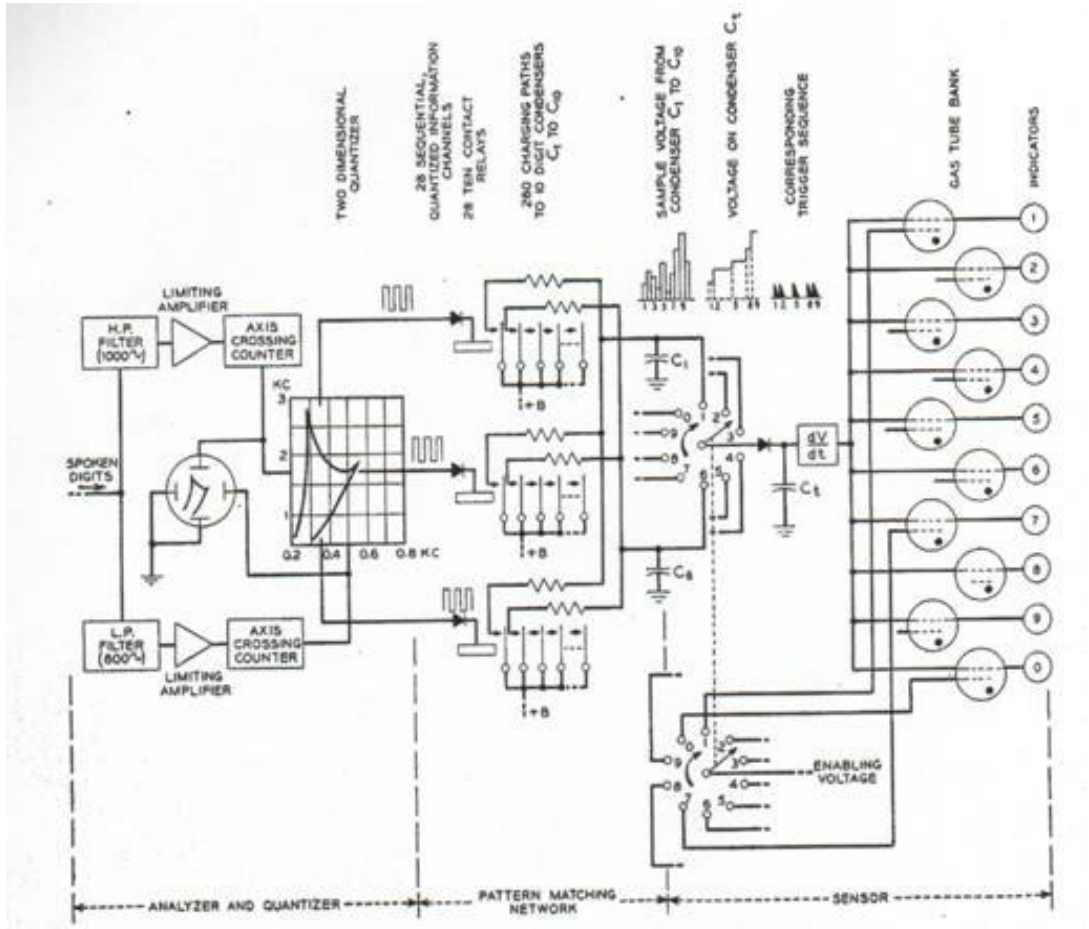
深度学习

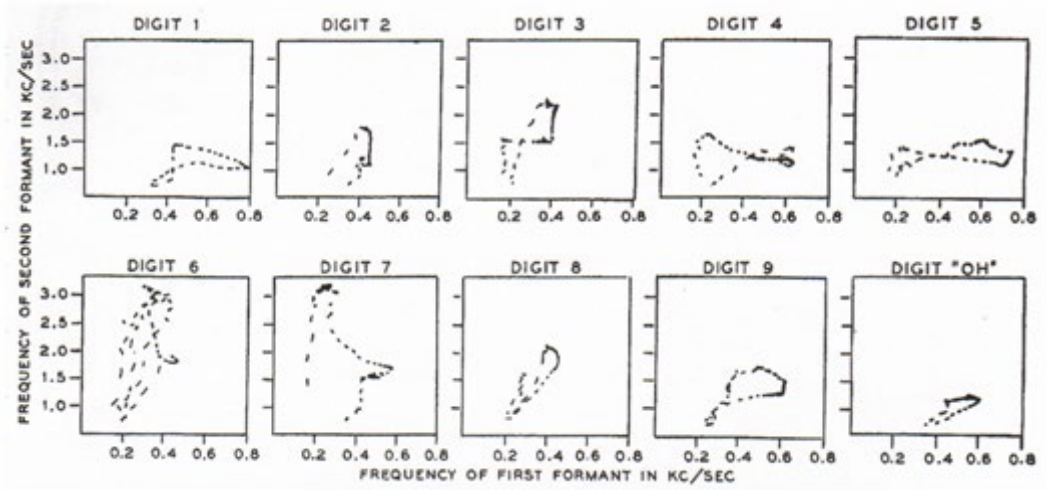
模板匹配 Audry识别10个英文数字	DTW VQ	HMM	GMM-HMM	DNN	DNN-HMM	End-to-End: CTC RNN-T Attention Transformer
1950s	1970s	1980S	1990S	2006	2012	2015以来

孤立词识别 → **大规模连续语音识别** → **复杂场景识别**

History of ASR

- 1952 – Automatic Digit Recognition (AUDREY)
  - Davis, Biddulph, Balashek (Bell Laboratories)





History of ASR

- 1960’ s – Speech Processing and Digital Computers
- AD/DA converters and digital computers start appearing in the labs

History of ASR

- 1969 – Whither Speech Recognition?

J. R. Pierce
Executive Director,
Bell Laboratories

General purpose speech recognition seems far away. Social-purpose speech recognition is severely limited. It would seem appropriate for people to ask themselves why they are working in the field and what they can expect to accomplish…

It would be too simple to say that work in speech recognition is carried out simply because one can get money for it. That is a necessary but not sufficient condition. We are safe in asserting that speech recognition is attractive to money. The attraction is perhaps similar to the attraction of schemes for turning water into gasoline, extracting gold from the sea, curing cancer, or going to the moon. One doesn't attract thoughtlessly given dollars by means of schemes for cutting the cost of soap by 10%. To sell suckers, one uses deceit and offers glamour…

Most recognizers behave, not like scientists, but like mad inventors or untrustworthy engineers. The typical recognizer gets it into his head that he can solve "the problem." The basis for this is either individual inspiration (the "mad inventor" source of knowledge) or acceptance of untested rules, schemes, or information (the untrustworthy engineer approach).

The Journal of the Acoustical Society of America, June 1969

History of ASR

- 1971-1976: The ARPA SUR project
- Despite anti-speech recognition campaign led by Pierce Commission ARPA launches 5 year Spoken Understanding Research program
- Goal: 1000-word vocabulary, 90% understanding rate, near real time on 100 mips machine
- 4 Systems built by the end of the program
  - SDC (24%)
  - BBN’ s HWIM (44%)
  - CMU’ s Hearsay II (74%)
  - CMU’ s HARPY (95% -- but 80 times real time!)
- Rule-based systems except for Harpy
  - Engineering approach: search network of all the possible utterances

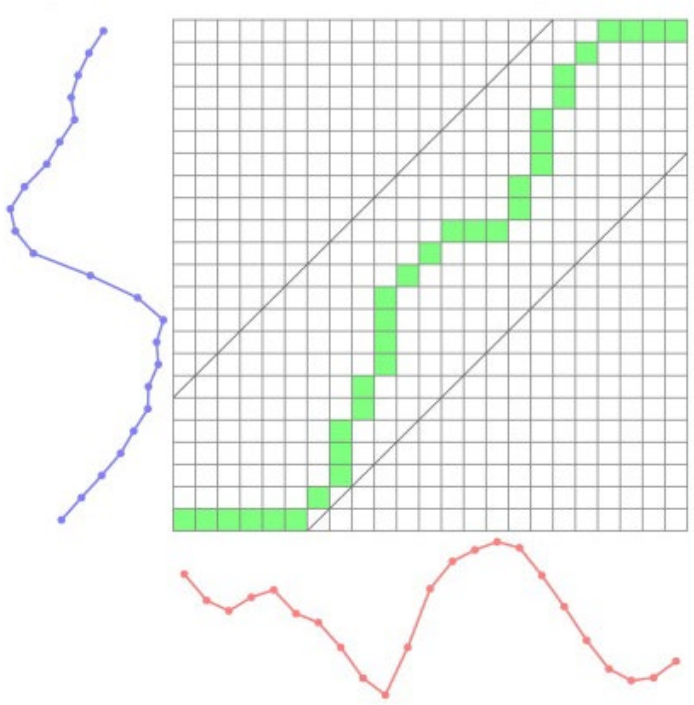
History of ASR

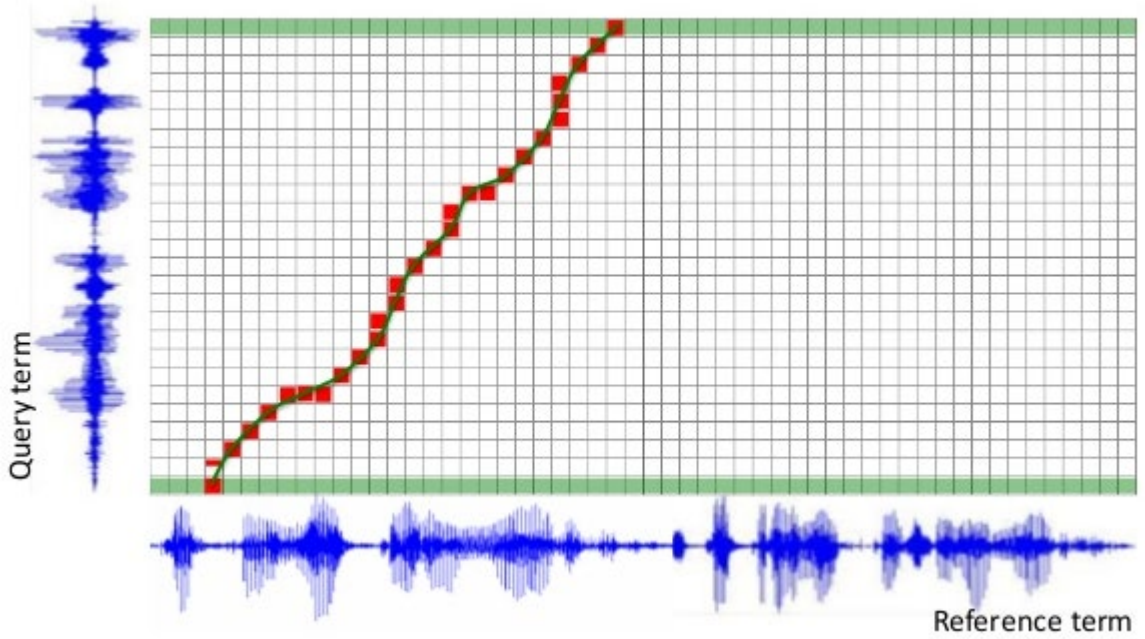
- 1971-1976: The ARPA SUR project
- Lack of clear evaluation criteria
  - ARPA felt systems had failed
  - Project not extended
- Speech Understanding: too early for its time
- Need a standard evaluation method

History of ASR

- 1970’ s – Dynamic Time Warping (DTW)
  - The Brute Force of the Engineering Approach

◆ 比较序列相似性　◆ 语音不等长匹配　◆ 小词汇量孤立词识别　◆ 特定人识别



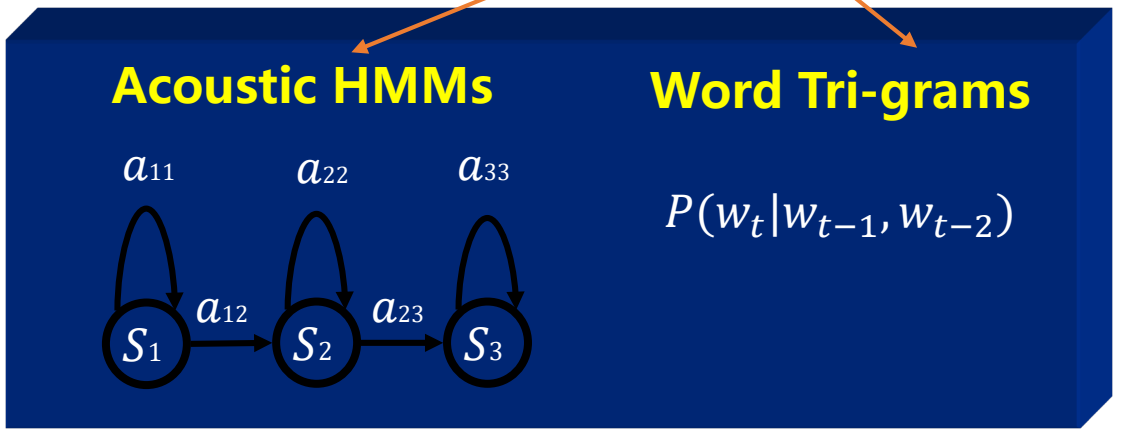


History of ASR

- 1980s -- The Statistical Approach
- Based on work on Hidden Markov Models done by Leonard Baum at IDA, Princeton in the late 1960s
- Purely statistical approach pursued by Fred Jelinek and Jim Baker, IBM T.J.Watson Research
- Foundations of modern speech recognition engines

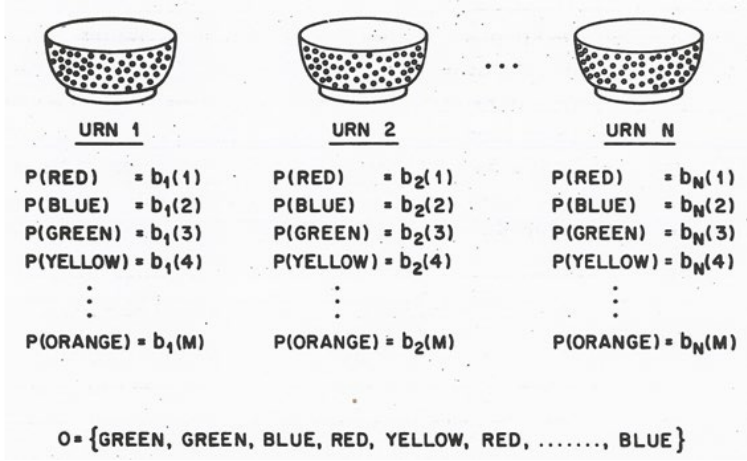
Fred Jelinek

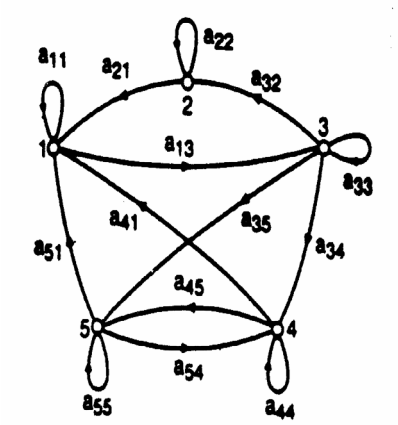
$$\hat{W} = \arg\max_{W} P(A|W)P(W)$$



History of ASR

- 1980-1990 – Statistical approach becomes ubiquitous
- Lawrence Rabiner, *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,* Proceeding of the IEEE, Vol. 77, No. 2, February 1989.





Markov Assumption:

$$P[q_t = j | q_{t-1} = i, q_{t-2} = k, \ldots] = P[q_t = j | q_{t-1} = i]$$

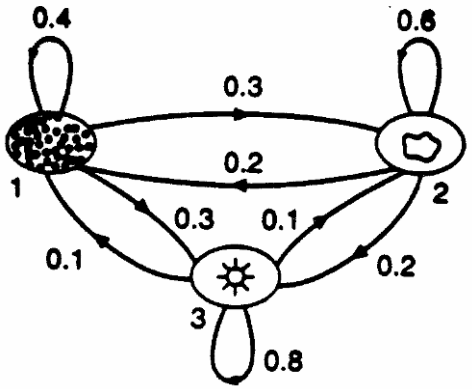
Set

$$a_{ij} = P[q_t = j | q_{t-1} = i] \quad 1 \le i, j \le N$$

Such that

$$a_{ij} \ge 0 \qquad \forall i, j$$

$$\sum_{j=1}^{N} a_{ij} = 1 \qquad \forall i$$

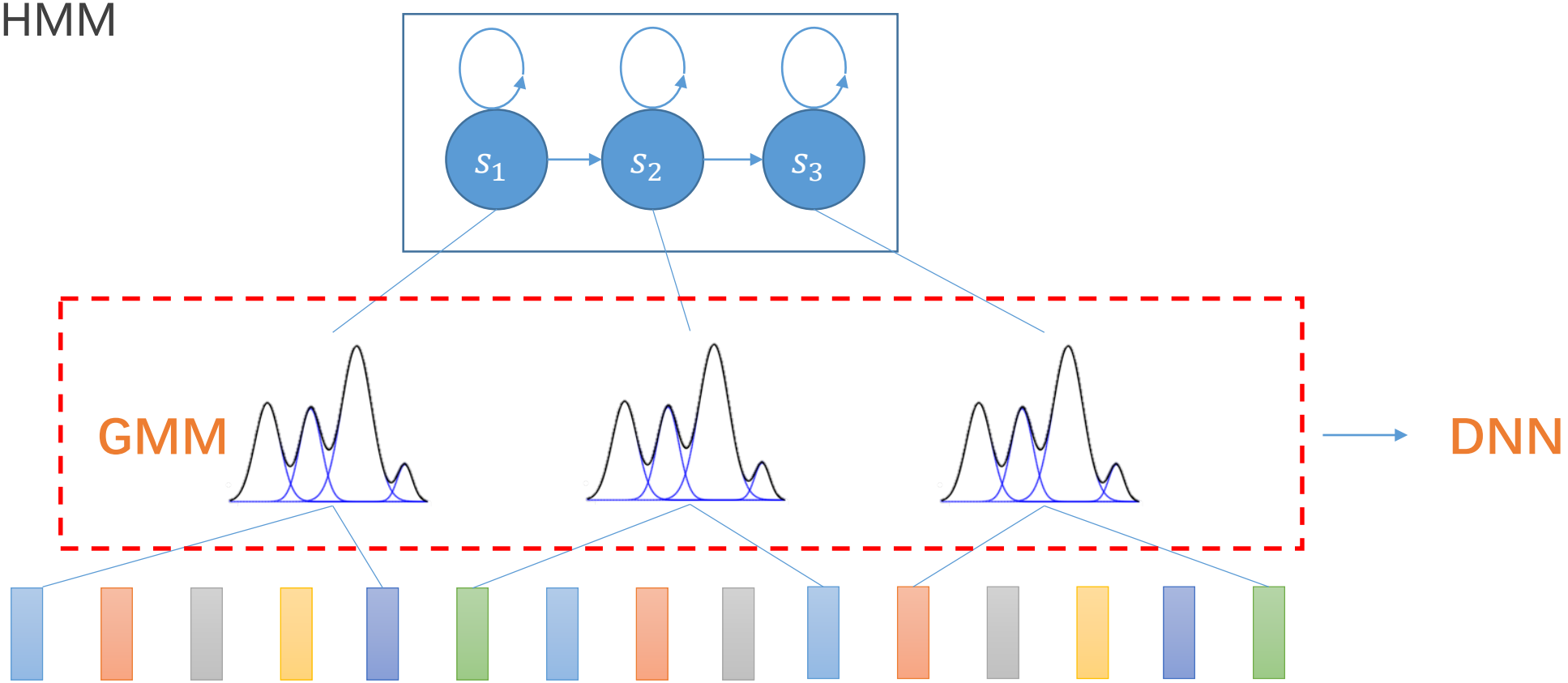


History of ASR

- Low noise conditions
- Large vocabulary
  - ~20,000-60,000 words or more…
- Speaker independent (vs speaker-dependent)
- Continuous speech (vs isolated-word)
- Multilingual, conversational
- World’ s best research systems:
  - Human-human speech: 5.5% Word Error Rate (WER)
  - Human-machine or monologue speech: ~3-5% WER

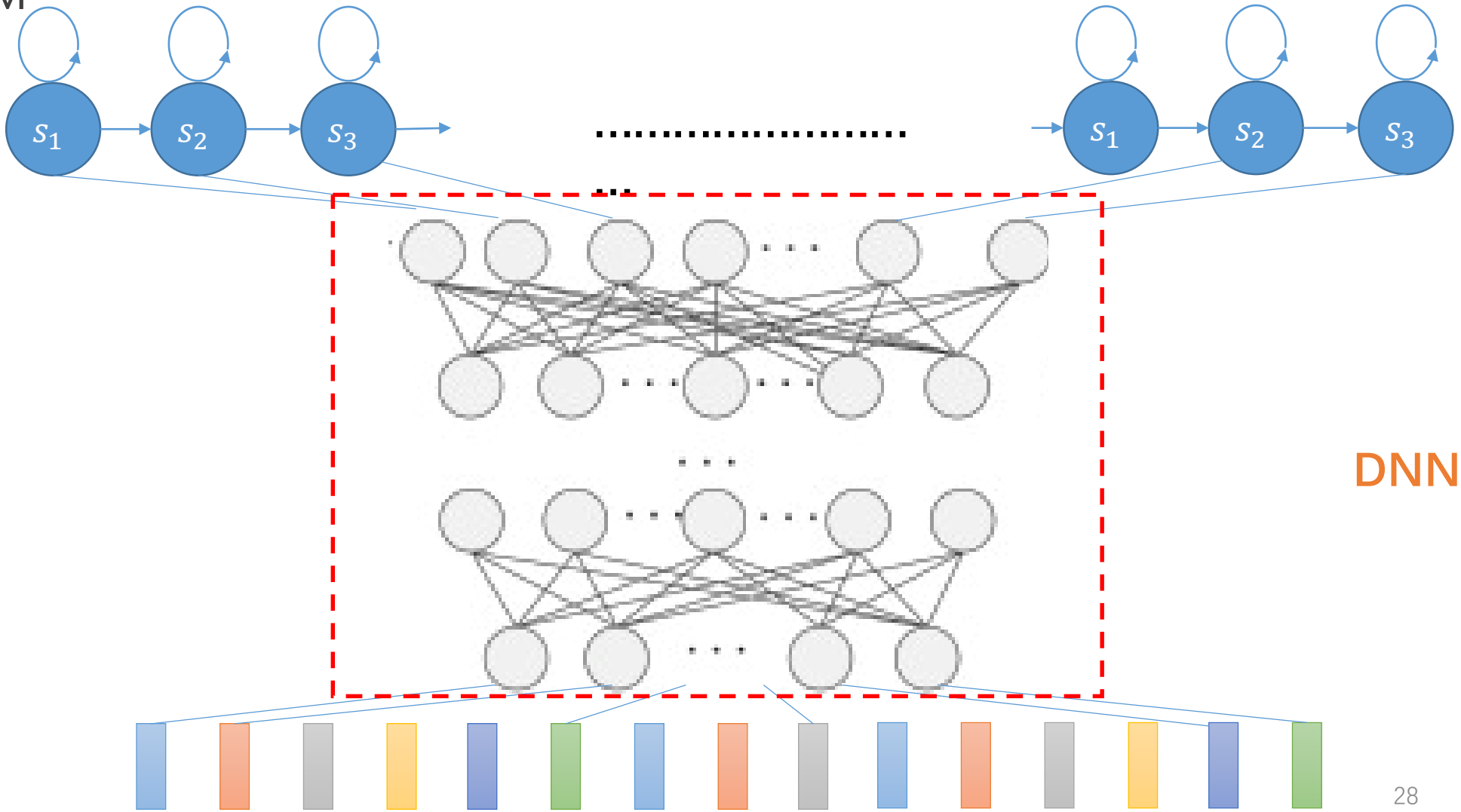
History of ASR

- DNN-HMM



History of ASR

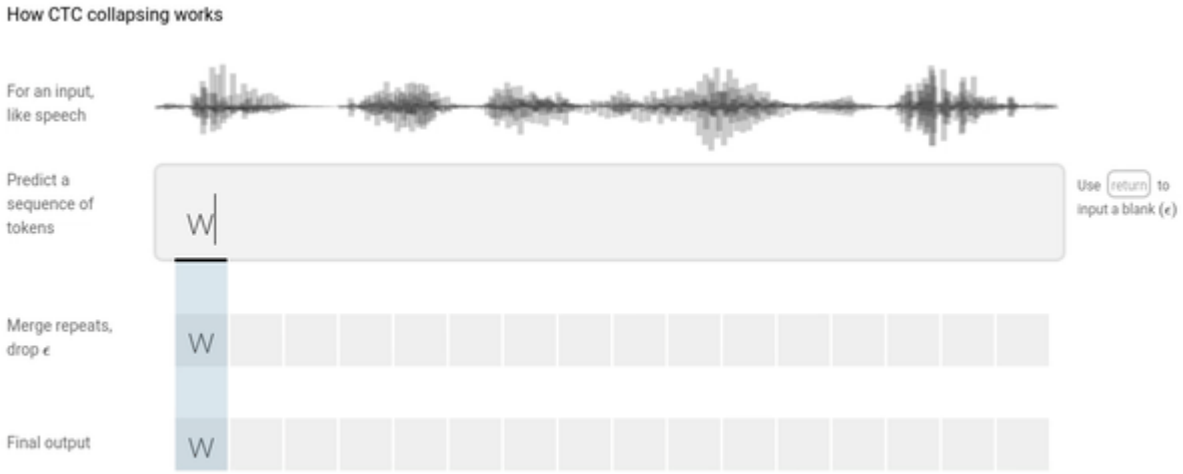
- DNN-HMM



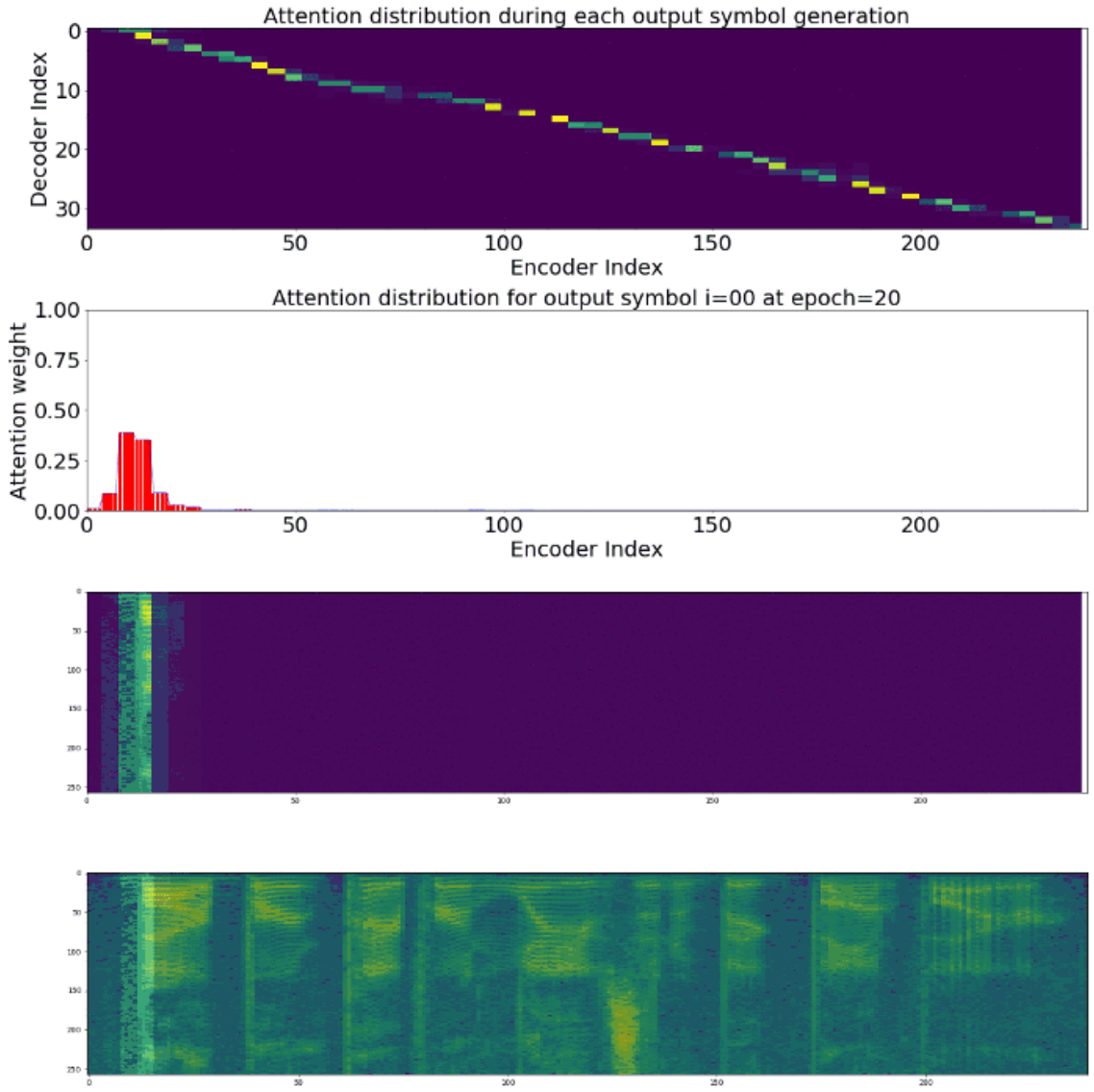
DNN模型

History of ASR

基于CTC损失函数的端到端模型



基于注意力机制的端到端模型



Why is speech recognition difficult?

Speaker	Tuned for a particular speaker, or speaker-independent? Adaptation to speaker characteristics
Environment	Noise, competing speakers, channel conditions (microphone, phone line, room acoustics)
Style	Continuously spoken or isolated? Planned monologue or spontaneous conversation?
Vocabulary	Machine-directed commands, scientific language, colloquial expressions
Accent/dialect	Recognize the speech of all speakers who speak a particular language
Other paralinguistics	Emotional state, social class, . . .
Language spoken	Estimated 7,000 languages, most with limited training resources; code-switching; language change

From a machine learning perspective

- As a classification problem: very high dimensional output space
- As a sequence-to-sequence problem: very long input sequence (although limited re-ordering between acoustic and word sequences)
- Data is often noisy, with many "nuisance" factors of variation in the data
- Very limited quantities of training data available (in terms of words) compared to text-based NLP
  - Manual speech transcription is very expensive (10x real time)
- Hierarchical and compositional nature of speech production and comprehension makes it difficult to handle with a single model

目 录

Open source tools

同济大学

HTK

- 剑桥大学Steve Young等人开发
- 2015年新版本实现了DNN-HMM

Kaldi

- 目前由Dan Povey博士维护
- C++

Espnet

- 一个端到端的语音处理工具集
- Python

Wenet-e2e

- 出门问问语音团队联合西工大语音实验室开发的开源语音识别工具包
- Python

目 录

Datasets

同济大学

TIMIT

- 英文语音识别库
- 630人，每人10句
- 比较简单

LibriSpeech

- 英文语音识别库
- 1000小时
- 朗读语音

Thchs-30

- 中文语音识别库
- 30小时
- 清华大学提供

AISHELL

- 中文语音识别库
- 400人，178小时
- 中国不同地区口音

目 录

Application in ASR

语音识别当中的实际问题

- 说话人自适应
- 噪声对抗与环境复杂性
- 新词处理与领域泛化
- 小语种识别
- 关键词唤醒与嵌入式系统

前沿课题

- 说话人识别
- 语种识别
- 情绪识别

1：基于融合特征的抑郁症检测

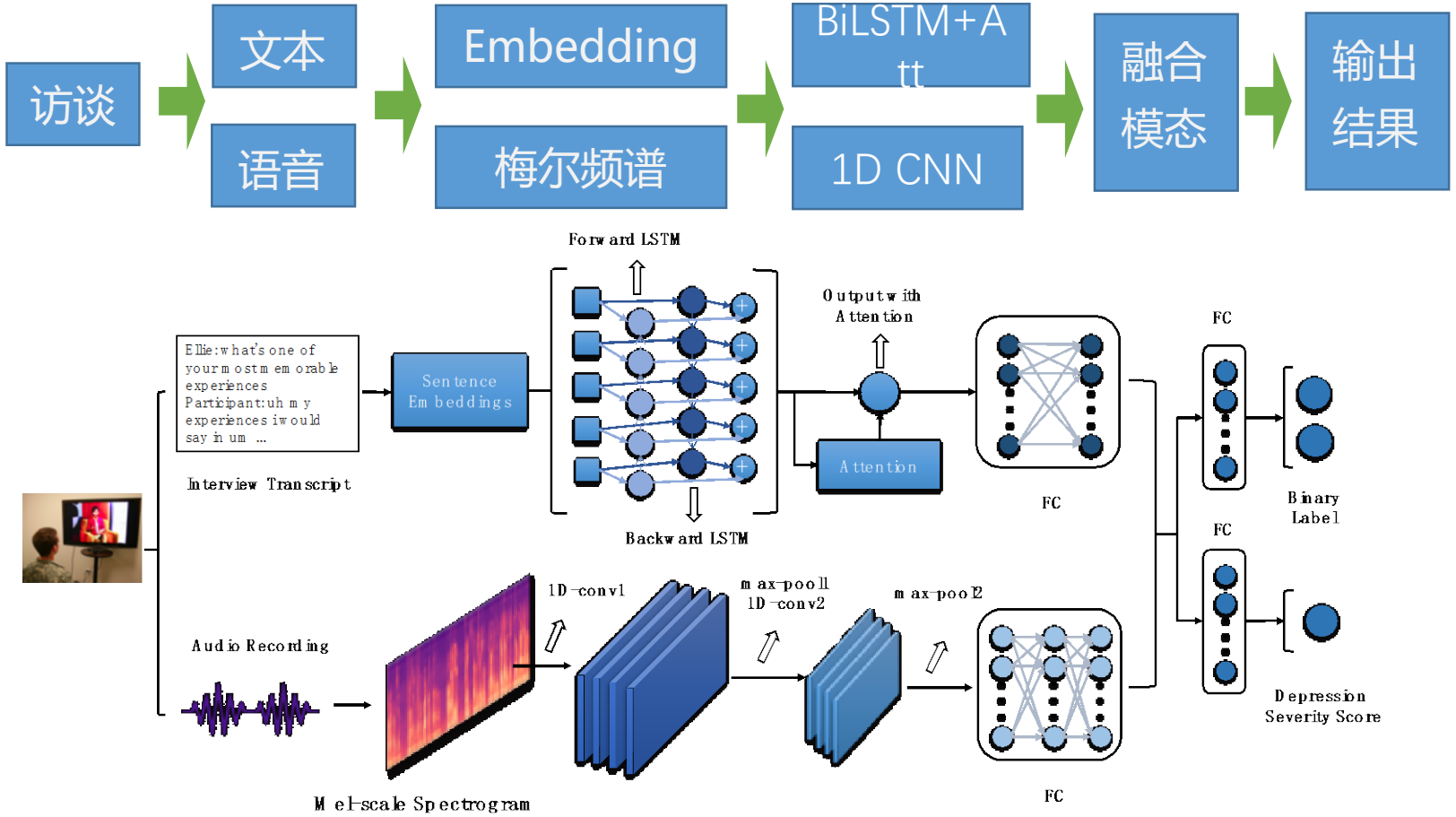
同济大学

抑郁症检测

模拟临床医生的诊断方式

结合语言，声音动作分析抑郁情况

基于融合特征的抑郁症检测模型结构



Ying Shen, Huiyu Yang, Lin Lin, "Automatic depression detection: An Emotional Audio-Textual Corpus and a GRU/BiLSTM-based model", in Proc. ICASSP 2022

1：基于融合特征的抑郁症检测

算法性能度量（DAIC）

音频模态: F1, Recall, MAE, RMSE

多模态融合: Recall, MAE, RMSE

Features	Models	Classification			Regression	
		F1 Score	Recall	Precision	MAE	RMSE
Audio	CNN-Augm [16]	0.67	0.58	0.78	-	-
	Williamson *et al.* [26]	0.50	-	-	5.36	6.74
	Ma *et al.* [17]	0.52	1.00	0.35	-	-
	Alhanai *et al.* [3]	0.63	0.56	0.71	5.13	6.50
	Yang *et al.* [28]	-	-	-	4.63	5.52
	Haque *et al.* [14]	-	-	-	5.78	-
	Proposed 1D CNN model	**0.81**	**0.92**	0.73	**4.25**	**5.45**
Text	Trf-Augm [16]	0.78	0.75	0.82	-	-
	Dinkel *et al.* [10]	**0.87**	0.83	0.93	-	-
	Haque *et al.* [14]	-	-	-	6.57	-
	Alhanai *et al.* [3]	0.67	0.80	0.57	5.18	6.38
	Sun *et al.* [22]	0.55	0.89	0.40	3.87	4.98
	Williamson *et al.* [26]	0.84	-	-	**3.34**	**4.46**
	Proposed BiLSTM model with an attention layer	**0.83**	0.83	0.83	**3.88**	**5.44**
Audio & Text	Trf+CNN-Augm [16]	**0.87**	0.83	0.91	-	-
	Alhanai *et al.* [3]	0.71	0.83	0.77	5.10	6.37
	Proposed multi-modal fusion network	**0.85**	**0.92**	0.79	**3.75**	**5.44**

算法性能度量（AViD）

AViD数据集仅提供音频

音频模态: MAE, RMSE

Models	Regression	
	MAE	RMSE
Baseline [16]	10.03	12.57
1D CNN	**9.30**	**11.55**

Ying Shen, Huiyu Yang, Lin Lin, "Automatic depression detection: An Emotional Audio-Textual Corpus and a GRU/BiLSTM-based model", in Proc. ICASSP 2022

2：基于流的声码器SiD-WaveFlow

基于流的声码器

将梅尔谱映射成waveform，得到音频

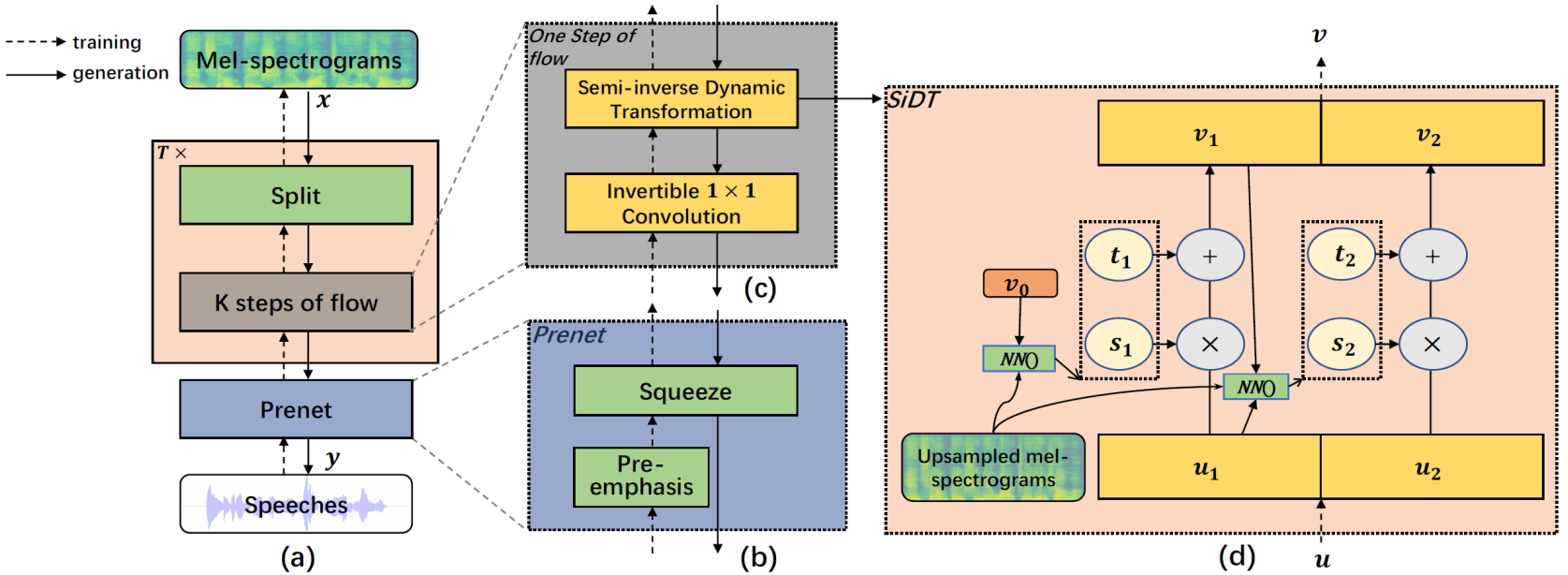


Figure 1: *The multi-scale architecture of SiD-WaveFlow. (a) SiD-WaveFlow transforms speeches to Mel-spectrograms at the training stage (dotted lines) and generates speeches from Mel-spectrograms (solid lines) at the speech generation stage; (b) The components in Prenet module; (c) The components in one step of flow; (d) The data flow in SiDT during the model training.*

Ying Shen, Huiyu Yang, Lin Lin, "Automatic depression detection: An Emotional Audio-Textual Corpus and a GRU/BiLSTM-based model", in Proc. ICASSP 2022

2：基于流的声码器SiD-WaveFlow

基于流的声码器

将梅尔谱映射成waveform，得到音频

Table 2: *Average MOS of synthesized speeches over 45 rounds of evaluation*

	CSMSC	LJ speech dataset
Ground Truth	3.754±0.007	3.067±0.002
Griffin-Lim	3.146±0.009	2.882±0.005
WaveGlow	3.324±0.001	2.909±0.001
Wave-IDLT	2.785±0.001	2.785±0.001
SiD-WaveFlow	**3.416±0.001**	**2.968±0.001**



Table 4: *The number of samples generated per second*

Models	Workstation	Raspberry Pi	Parameter#
WaveGlow	405k	4.4k	87.9M
Wave-IDLT	139k	failed	169.2M
SiD-WaveFlow	**522k**	**5.1k**	**63.1M**

Yuhan Li, Ying Shen*, Dongqing Wang and Lin Zhang, "SiD-WaveFlow: A Low-Resource Vocoder Independent of Prior Knowledge", in Proc. Interspeech 2022

3：低信噪比下的语音识别应答系统

低信噪比下语音识别应答系统

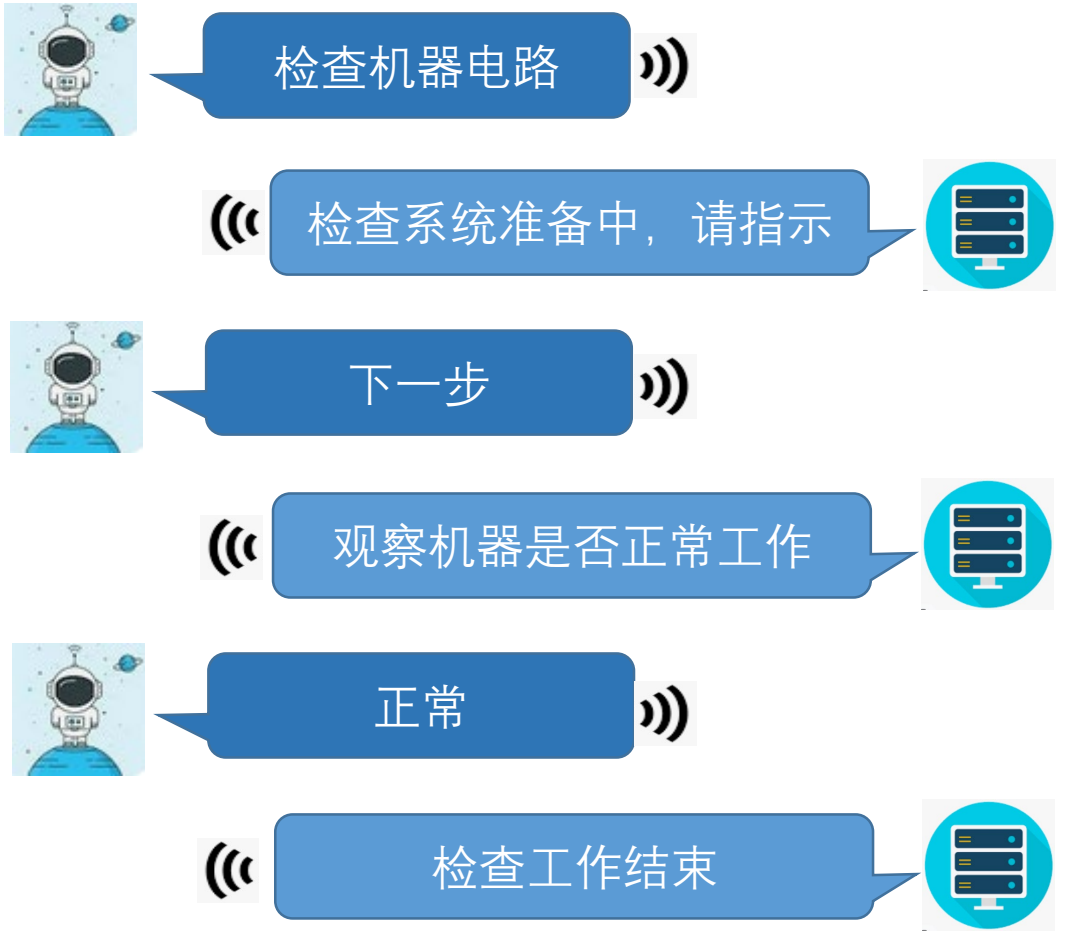
在空间站帮助宇航员发布命令，查找工作手册，获得操作指导。包括离线和在线子系统

工作流程

语音识别 – 策略应答 – 语音合成

其他需求

实时性
应答策略的可扩展性
合成语音的质量



3：低信噪比下的语音识别应答系统

语音识别应答系统架构

离线语音交互系统和在线语音交互系统

离线语音交互系统

DSP客户端

收集并发送语音

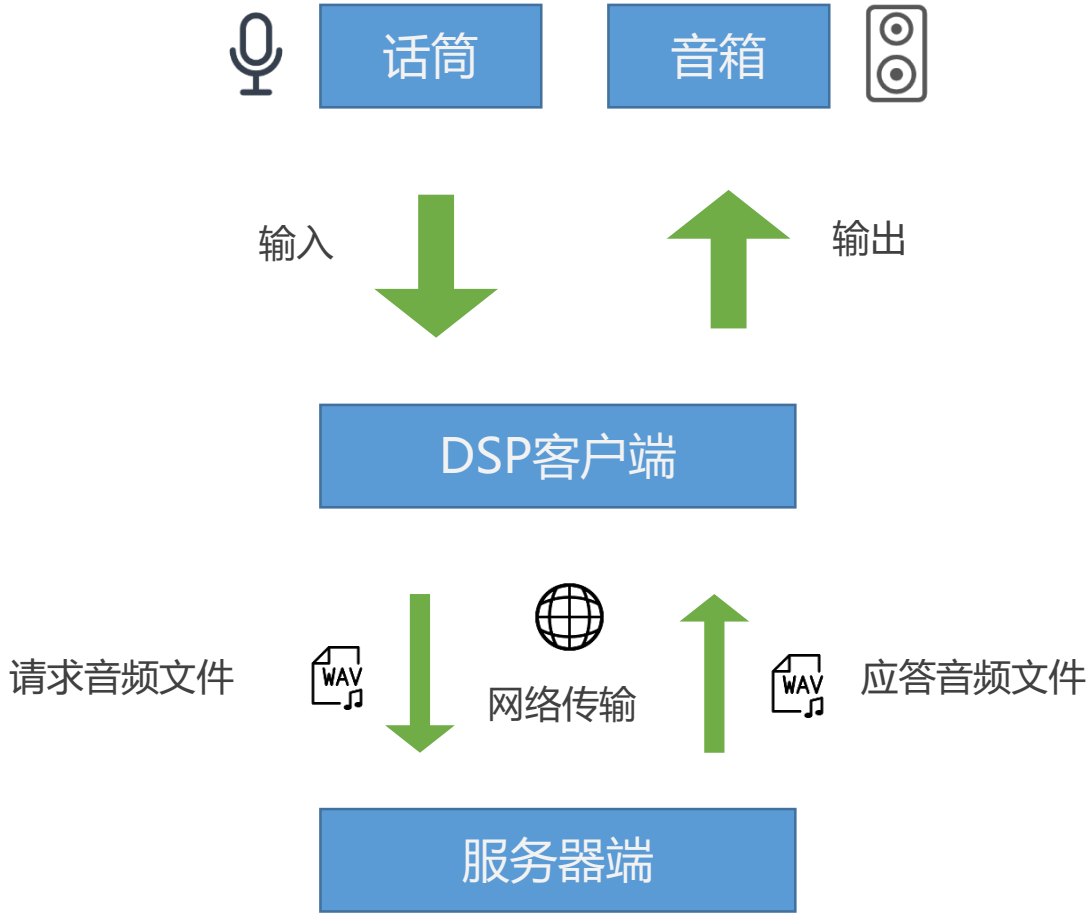
孤立词识别、简单应答、合成语音

在线语音交互系统

服务器端

请求音频文件、发送应答音频

连续语音识别、PyAIML、合成语音



3：低信噪比下的语音识别应答系统

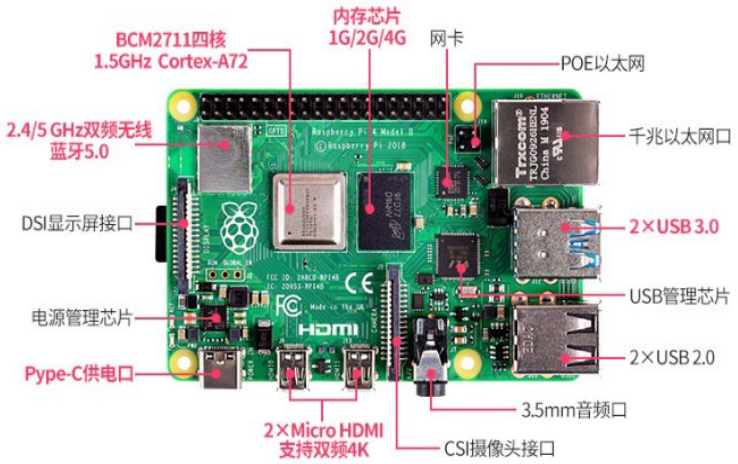
视频演示

4：基于树莓派的声纹识别系统

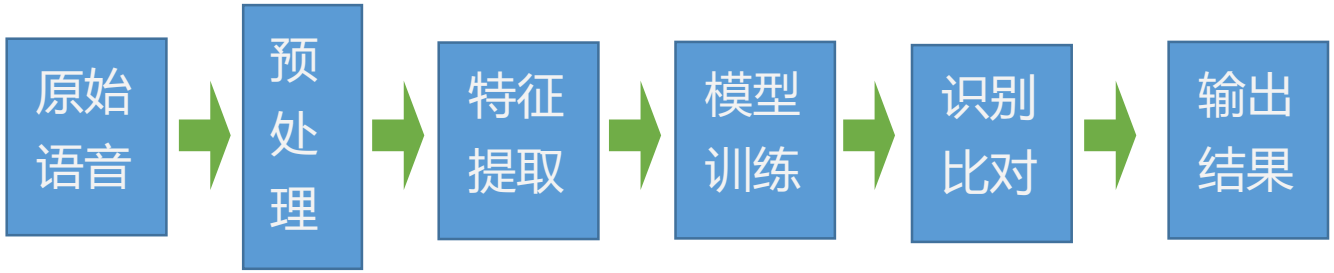
树莓派

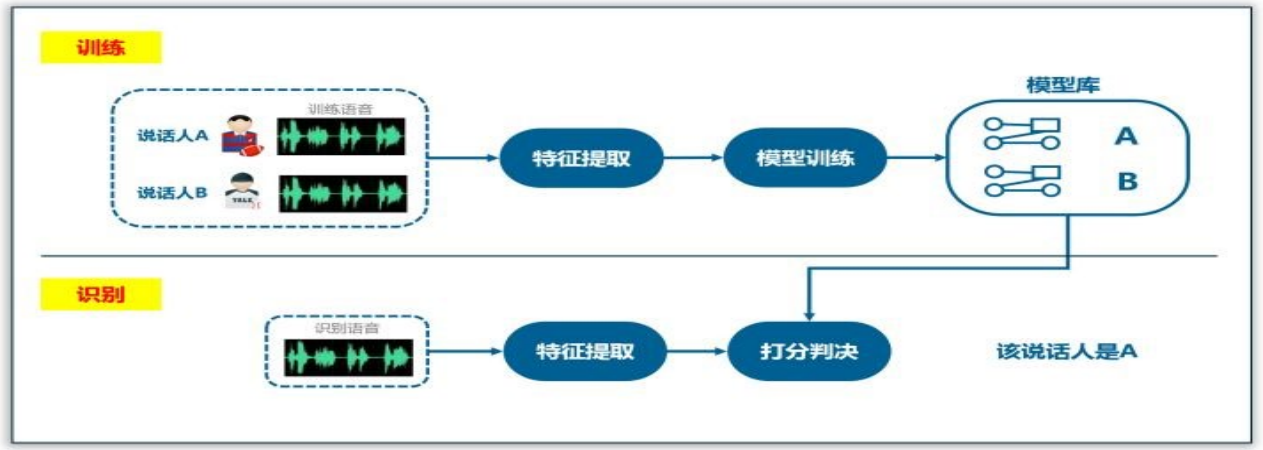
功能强大　　轻便小巧

可拓展性高



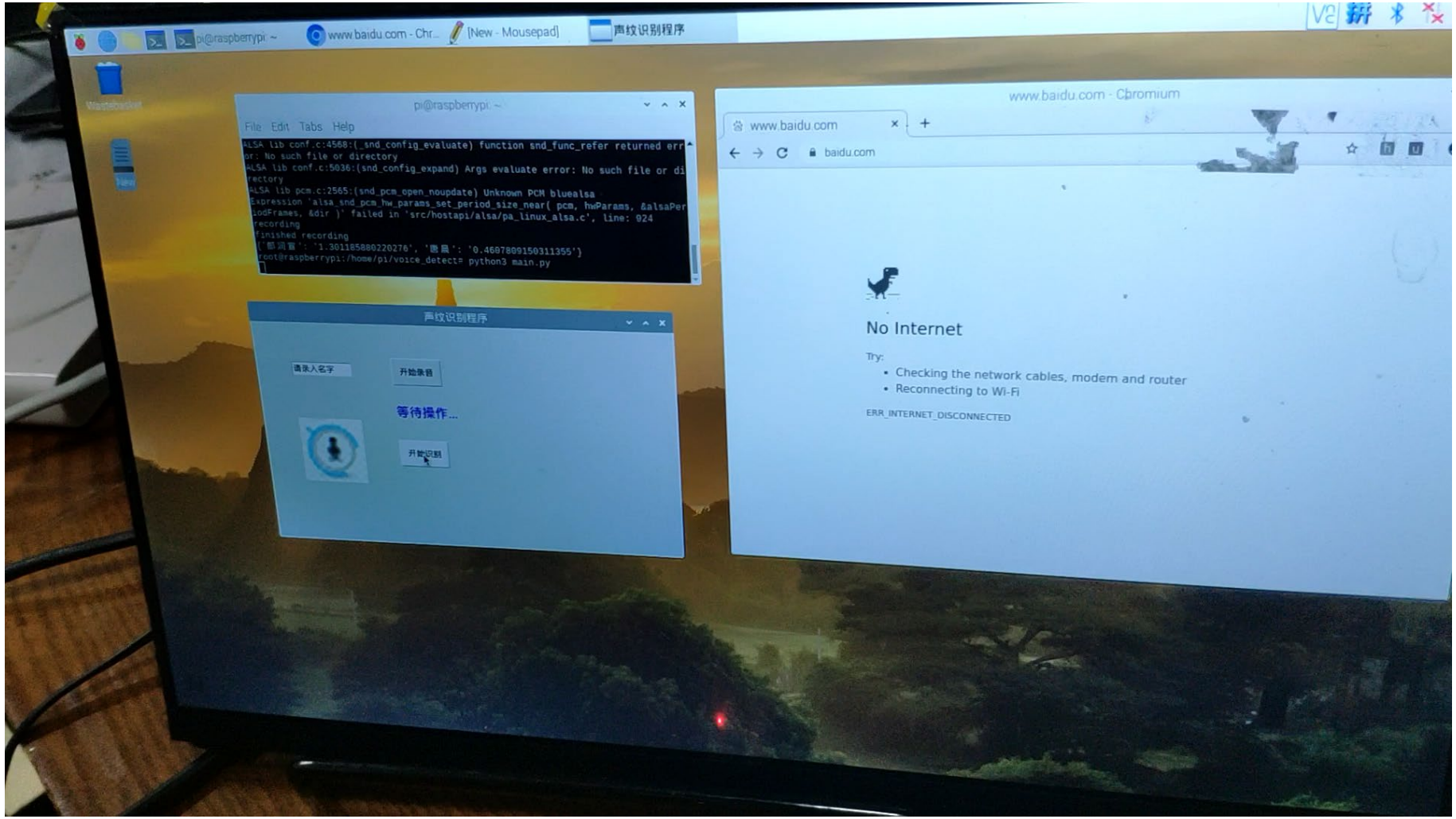
声纹识别工作流程





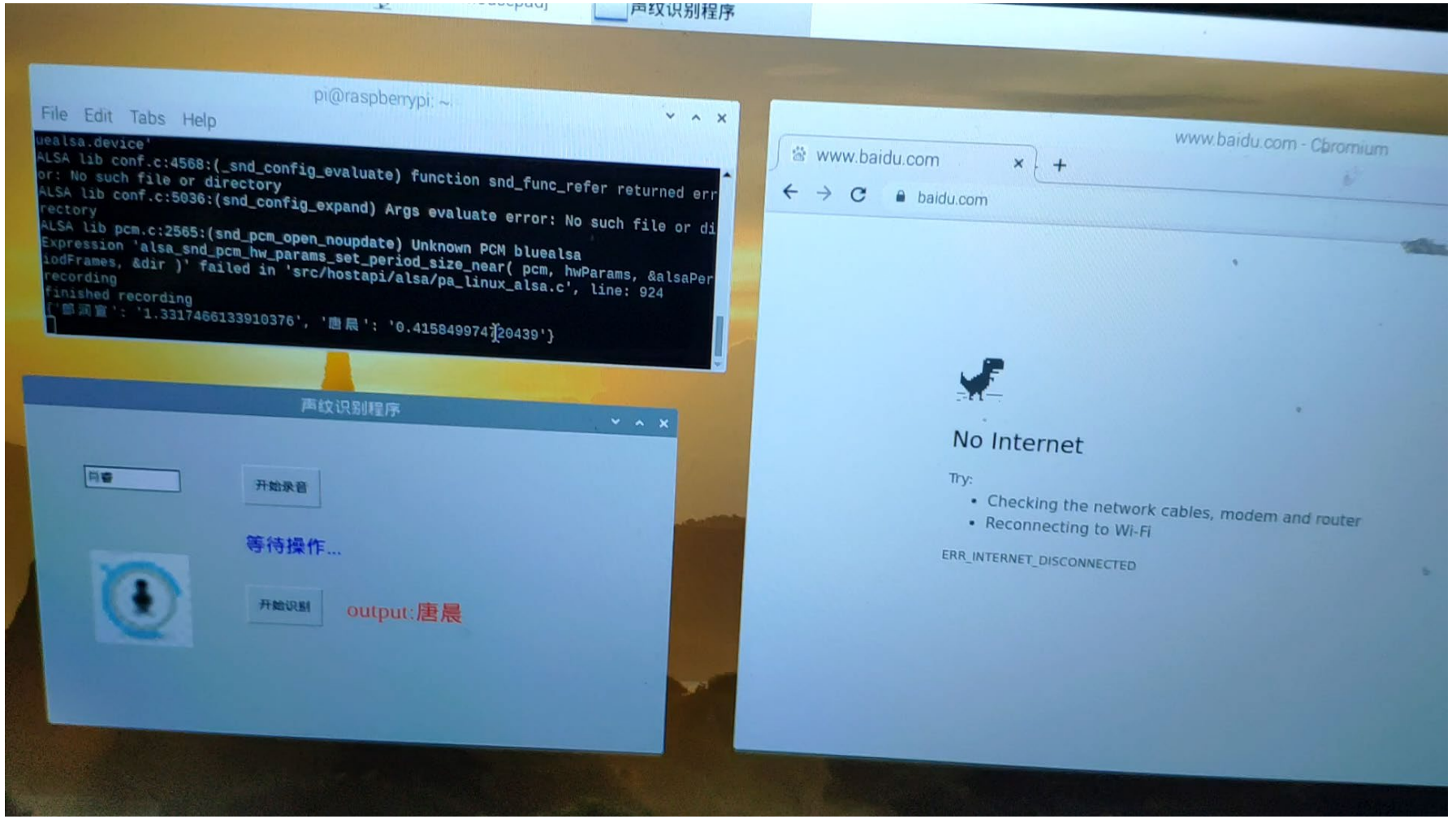
4：基于树莓派的声纹识别系统

实操视频演示



4：基于树莓派的声纹识别系统

实操视频演示



目 录

Existing problems

- 复杂场景
- 口语化
- 低资源
- 语种混杂