MATLAB EXPO

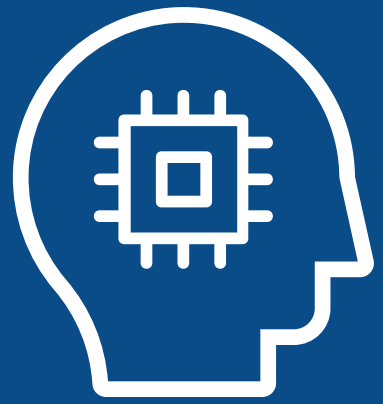
远不及此：AI识别语音关键词

马文辉

MathWorks®

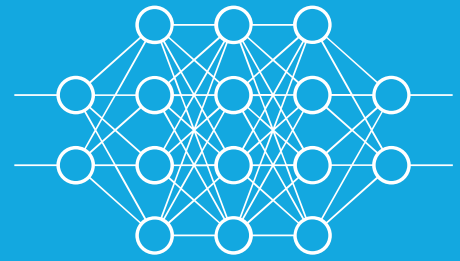
深度学习是推动人工智能发展的一项关键技术

人工智能
任何能使机器模仿人类智能的技术



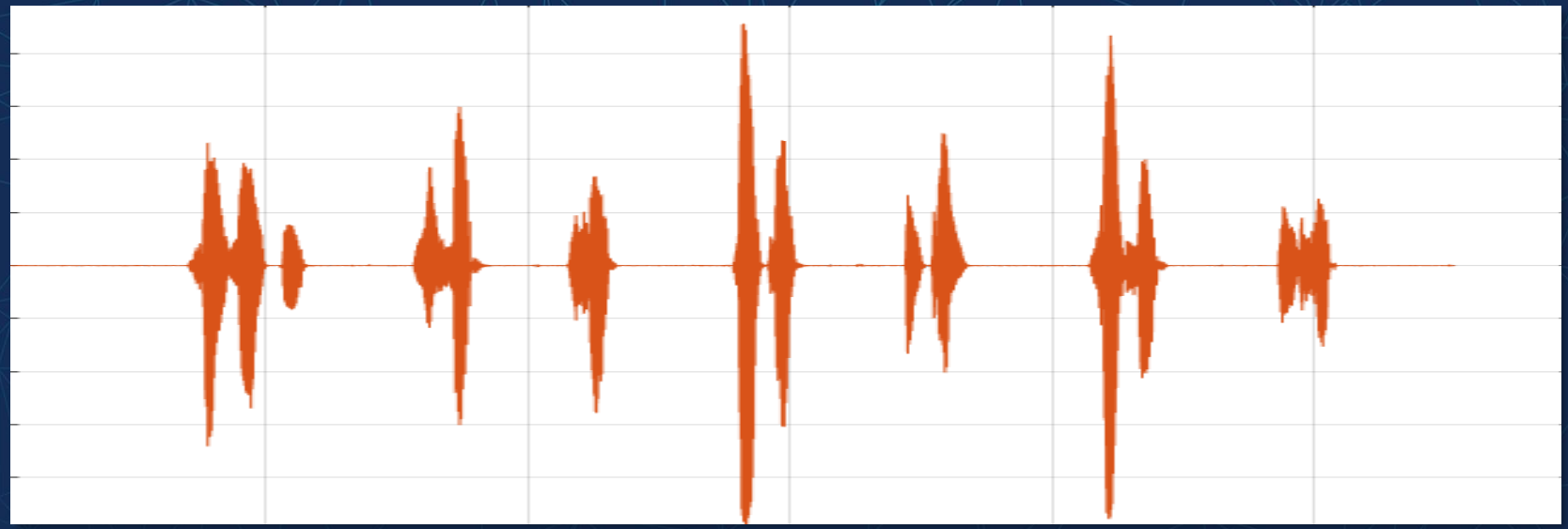
机器学习
使机器无需显式编程就能从数据中“学习”任务的方法

深度学习
具有多层的神经网络“直接”从数据中学习表示和任务的方法



1950s　　1980s　　2010s

为信号处理应用开发一个有效的深度学习应用需要什么？



实际案例：语音触发关键词

(基于云的语音助手)

回答: “设计有效的深度神经网络"

“双向长短期记忆网络层（LSTM, 每层带有150个隐藏单元）+ 全连接层(fullconnect) + softmax 层"

回答: “大量的数据，信号处理专业知识，以及针对特定应用的合适工具"

数据创建和访问

数据处理和转换

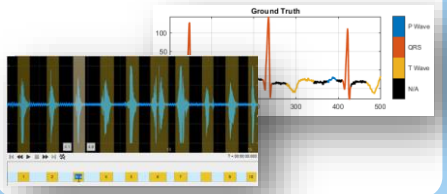
数据建模

模型部署

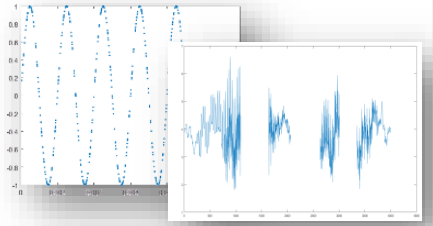
Data sources

Simulation and augmentation

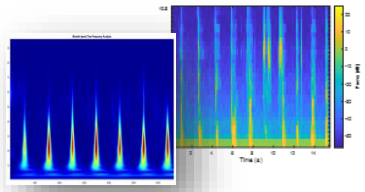
Data Labeling



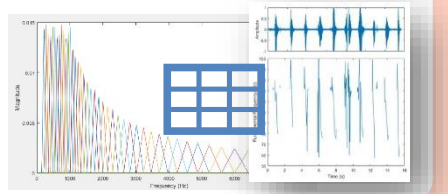
Pre-Processing



Transformation

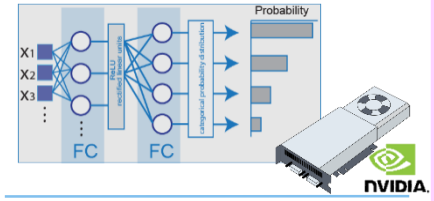


Feature extraction



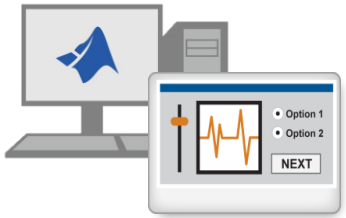
Import Reference Models/ Design from scratch

Hardware-Accelerated Training

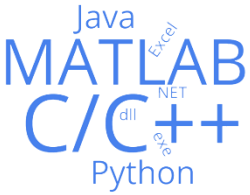


Analyze and tune hyperparameters

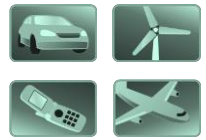
Desktop Apps



Enterprise Scale Systems

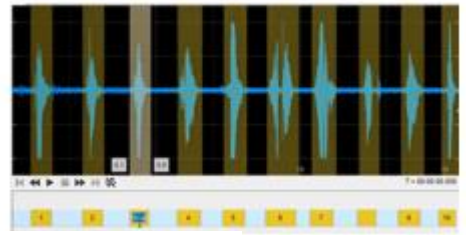


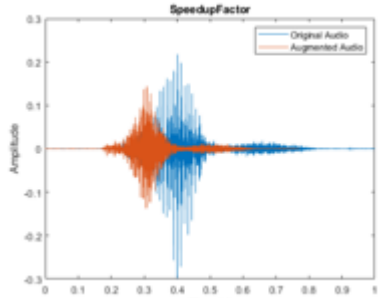
Embedded Devices and Hardware

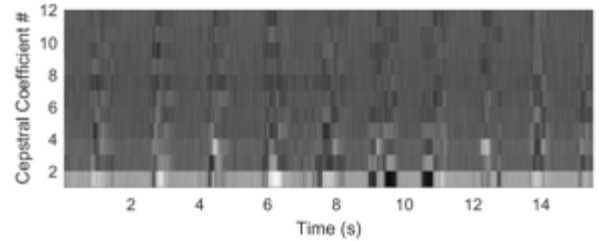


主要内容

- 设计深度神经网络
- 数据标注
- 数据增强 – 合成数据
- 数据转换
- 模型部署

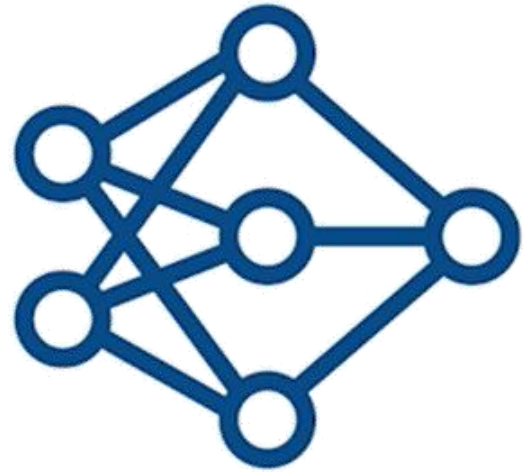


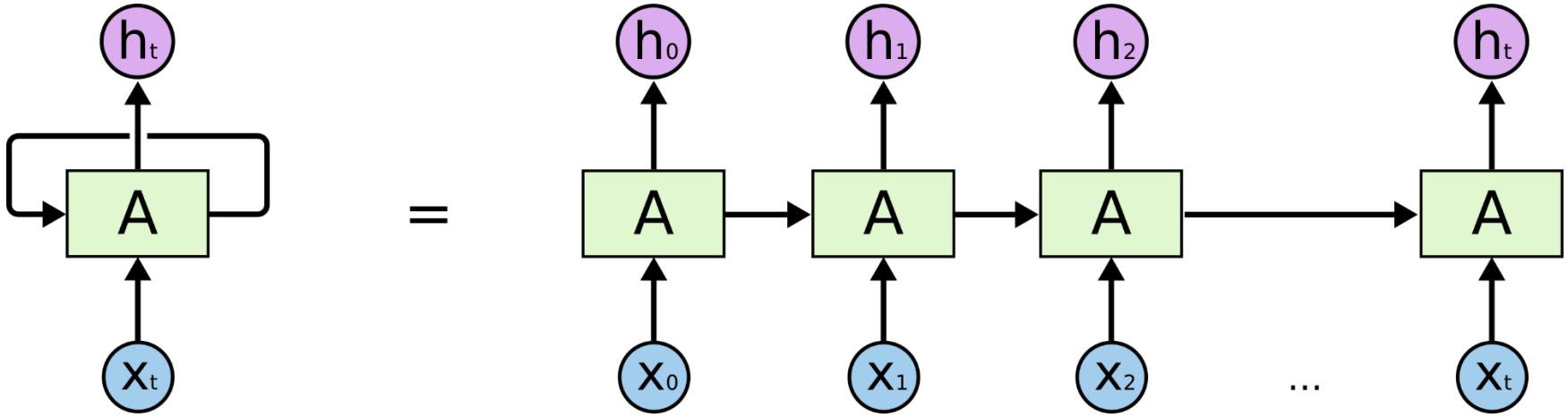




设计深度神经网络 - LSTM

```
layers = [ ...
    sequenceInputLayer(numFeatures)
    bilstmLayer(150,"OutputMode","sequence")
    bilstmLayer(150,"OutputMode","sequence")
    fullyConnectedLayer(2)
    softmaxLayer
    classificationLayer
    ];
```



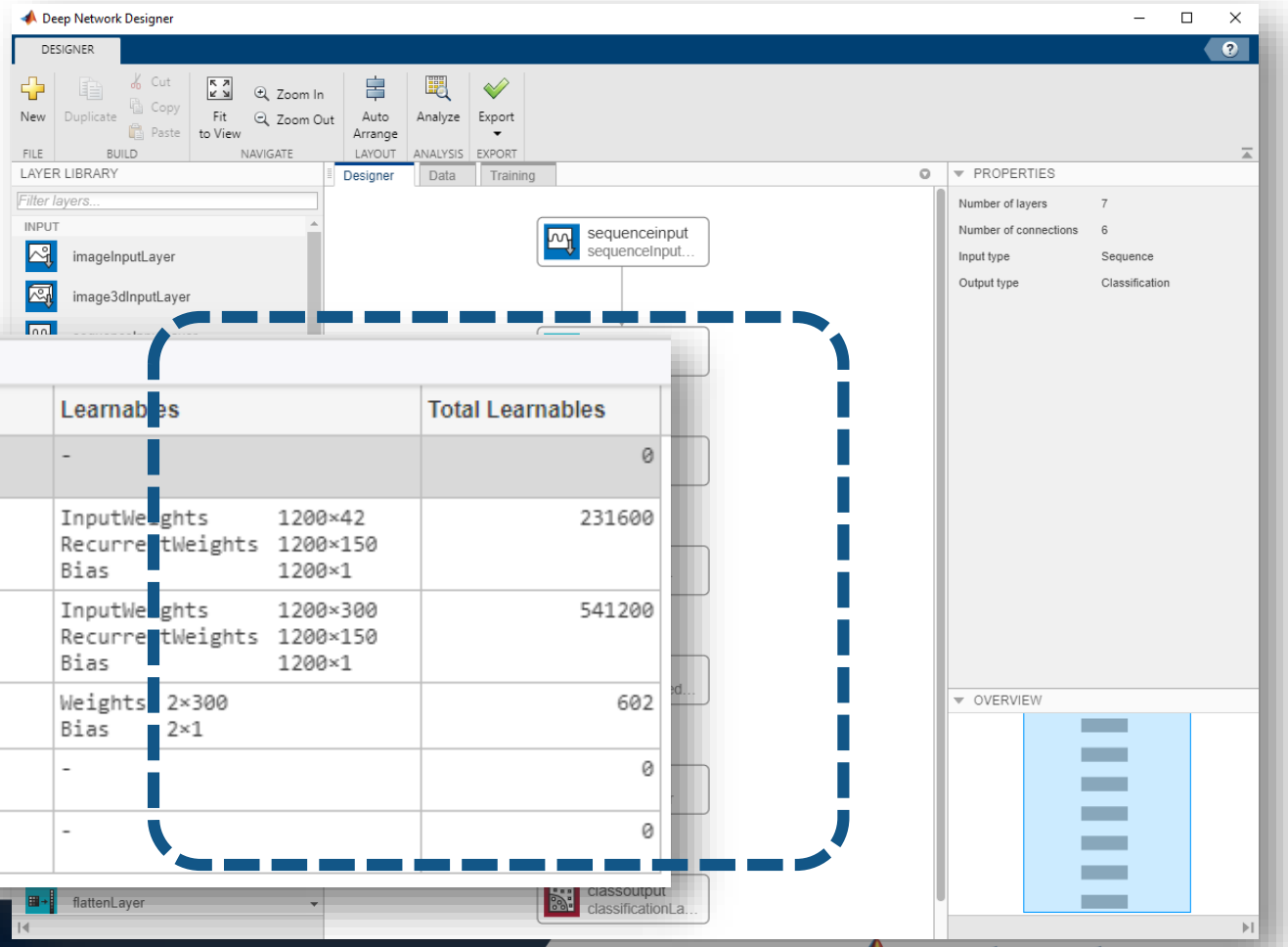


循环神经网络(Recurrent Neural Networks, **RNN**)

长短期记忆（Long Short Term Memory，**LSTM**) 网络

设计深度神经网络 - LSTM

```
layers = [ ...
    sequenceInputLayer(numFeatures)
    bilstmLayer(150,"OutputMode","sequence")
    bilstmLayer(150,"OutputMode","sequence")
    fullyConnectedLayer(2)
    softmaxLayer
    classificationLayer
    ];
```



ANALYSIS RESULT

	Name	Type	Activations	Learnables	Total Learnables
1	sequenceinput Sequence input with 42 dimensions	Sequence Input	42	-	0
2	biLSTM_1 BiLSTM with 150 hidden units	BiLSTM	300	InputWeights 1200×42 RecurrentWeights 1200×150 Bias 1200×1	231600
3	biLSTM_2 BiLSTM with 150 hidden units	BiLSTM	300	InputWeights 1200×300 RecurrentWeights 1200×150 Bias 1200×1	541200
4	fc 2 fully connected layer	Fully Connected	2	Weights 2×300 Bias 2×1	602
5	softmax softmax	Softmax	2	-	0
6	classoutput crossentropyex	Classification Output	-	-	0

借助已有科研成果

Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling

Haşim Sak, Andrew Senior, Françoise Beaufays

Long short-term memory for speaker generalization in supervised speech separation

Jitong Chen[a)] and DeLiang Wang[b)]

Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio 43210, USA

d online 23

An Improved Residual LSTM Architecture for Acoustic Modeling

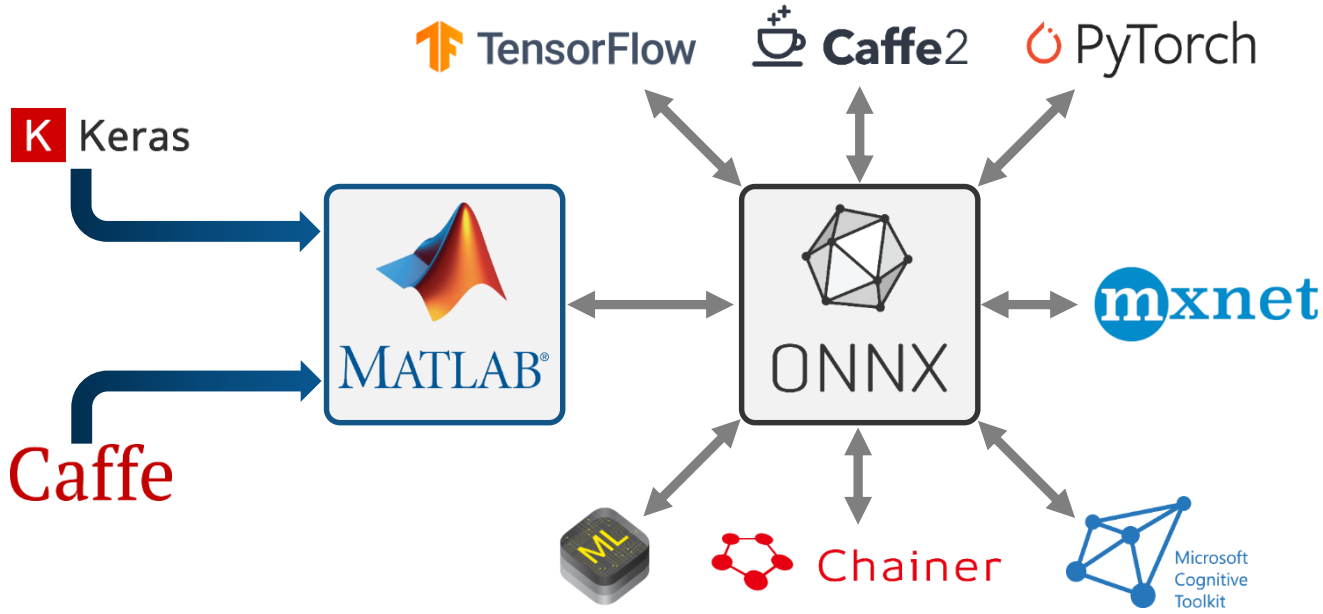
Lu Huang
Department of Electronic Engineering
Tsinghua University
Beijing, China
e-mail: huanglu.th@gmail.com

Jiasong Sun
Department of Electronic Engineering
Tsinghua University
Beijing, China
e-mail: sunjiasong@tsinghua.edu.cn

Ji Xu
b of Speech Acoustics & Content Understanding
ute of Acoustics, Chinese Academy of Sciences

Yi Yang
Department of Electronic Engineering
Tsinghua University

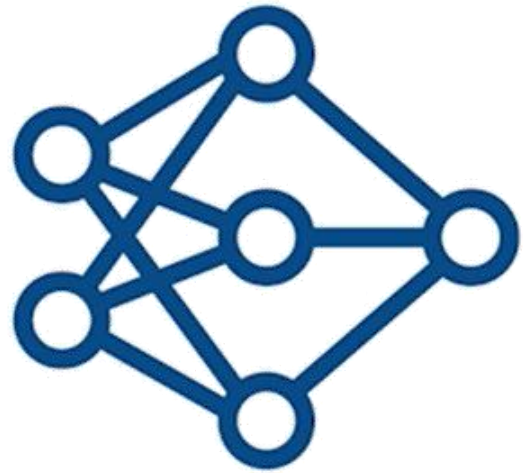
...或者导入其它平台已训练好的网络



ONNX – Open Neural Network Exchange

训练深度神经网络

```
layers = [ ...
    sequenceInputLayer(numFeatures)
    bilstmLayer(150,"OutputMode","sequence")
    bilstmLayer(150,"OutputMode","sequence")
    fullyConnectedLayer(2)
    softmaxLayer
    classificationLayer
    ];

maxEpochs     = 10;
miniBatchSize = 64;
options = trainingOptions("adam", ...
    "InitialLearnRate",1e-4,...
    "MaxEpochs",maxEpochs, ...
    "MiniBatchSize",miniBatchSize, ...
    "Shuffle","every-epoch",...
    "Verbose",false, ...
    "ValidationFrequency",floor(numel(TrainingFeatures)/miniBatchSize),...
    "ValidationData",{FeaturesValidationClean.',BaselineV},...
    "Plots","training-progress",...
    "LearnRateSchedule","piecewise",...
    "LearnRateDropFactor",0.1, ...
    "LearnRateDropPeriod",5);

[net,info] = trainNetwork(TrainingFeatures,TrainingMasks,layers,options);
```



b88a9ffecebd:1 () - TigerVNC

MATLAB R2019b

HOME PLOTS APPS EDITOR PUBLISH VIEW

Search Documentation

Andy

/ ▸ home ▸ matlab ▸ Documents ▸ AudioWebinar ▸ Code

Workspace

Name	Value
expectedNumPartitions	128
klstm	4
kovlp	4
loadFeatures	1
LSTMSize	150
LSTMSizes	[75,100,125,150]
LSTMSizes	[75,100,125,150]
M	*1x4 cell*
net	*1x1 SeriesNetwork*
netLayers	*6x1 Layer*

Current Folder | Command Window

fx >>

Editor - TrainSingleNetwork.m

Variables - feat

TrainSingleNetwork.m

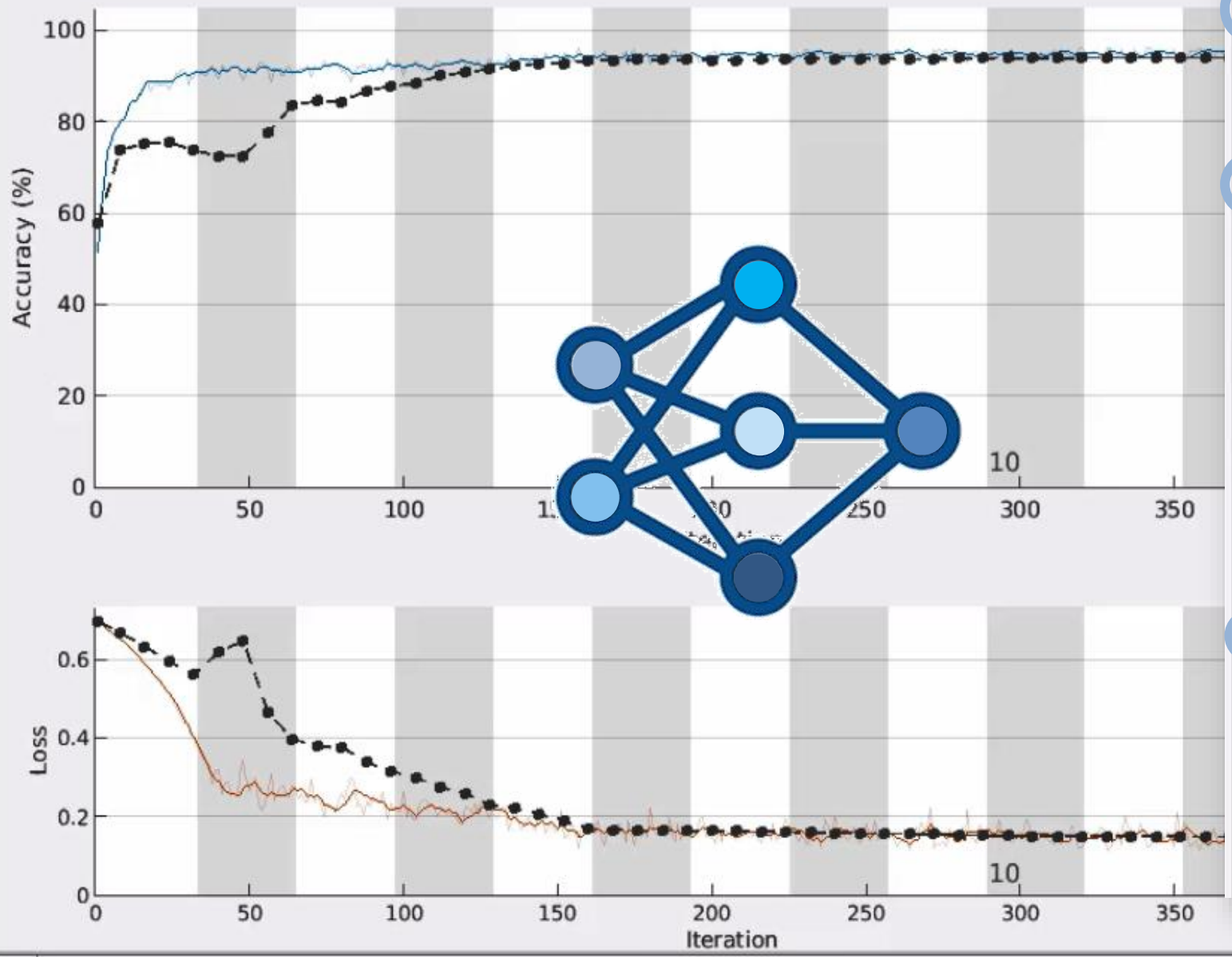
```
40 -    netLayers = [ ...
41          sequenceInputLayer(numFeatures)
42          bilstmLayer(LSTMSizes(klstm),"OutputMode","sequence")
43          bilstmLayer(LSTMSizes(klstm),"OutputMode","sequence")
44          fullyConnectedLayer(2)
45          softmaxLayer
46          classificationLayer
47          ];
48
49 -    trainOptions = trainingOptions("adam", ...
50          "InitialLearnRate",1e-4, ...
51          "MaxEpochs",12, ...
52          "MiniBatchSize",4, ...
53          "Shuffle","every-epoch", ...
54          "Verbose",false, ...
55          "ValidationFrequency",8, ...
56          "ValidationData",{ValidationFeatures{kovlp},ValidationMasks{kovlp}}, ...
57          "Plots","training-progress", ...
58          "LearnRateSchedule","piecewise", ...
59          "LearnRateDropFactor",0.1, ...
60          "LearnRateDropPeriod",5,...
61          "SequenceLength","Shortest");
62
63      %% Network training
64
65 -    tic;
66 -    net = trainNetwork(trainingFeatures,trainingMasks,netLayers,trainOptions);
67 -    fprintf('Training the network took %g s\n',toc);
68
69
```

Busy | script | Ln 63 Col 1

b88a9ffecebd:1 () - TigerVNC

Training Progress (20-Mar-2020 12:21:25)

Training Progress (20-Mar-2020 12:21:25)



Results

Validation accuracy:	93.96%
Training finished:	Reached final iteration

Training Time

Start time:	20-Mar-2020 12:21:25
Elapsed time:	8 min 15 sec

Training Cycle

Epoch:	12 of 12
Iteration:	384 of 384
Iterations per epoch:	32
Maximum iterations:	384

Validation

Frequency:	8 iterations
Patience:	Inf

Other Information

Hardware resource:	Single GPU
Learning rate schedule:	Piecewise

Accuracy

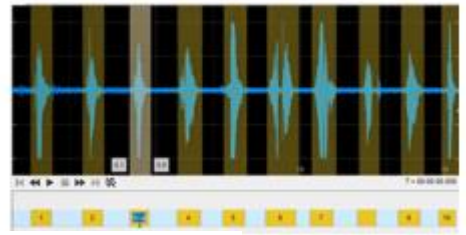
Training (smoothed)

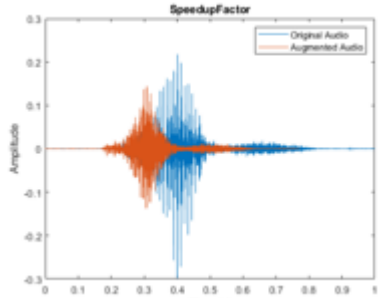
Training

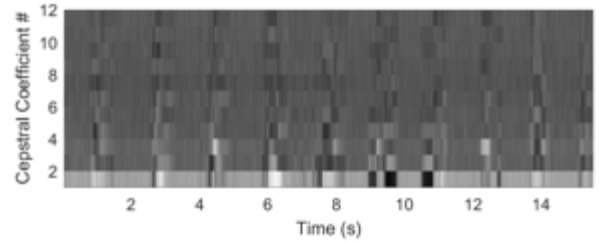
Validation

主要内容

- 深度神经网络
- 数据标注
- 数据生成 – 合成数据与数据增强
- 输入数据转换
- 模型部署

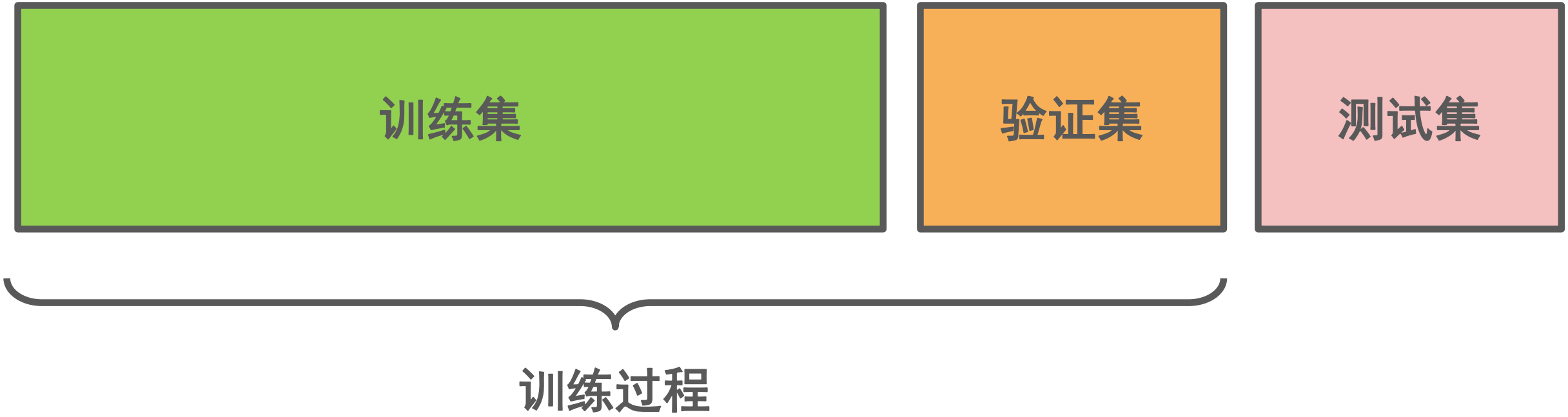






训练集, 验证集, 测试集

GB – TB



训练集, 验证集, 测试集

全数据集 (**data + labels**)

实际数据，正确标注

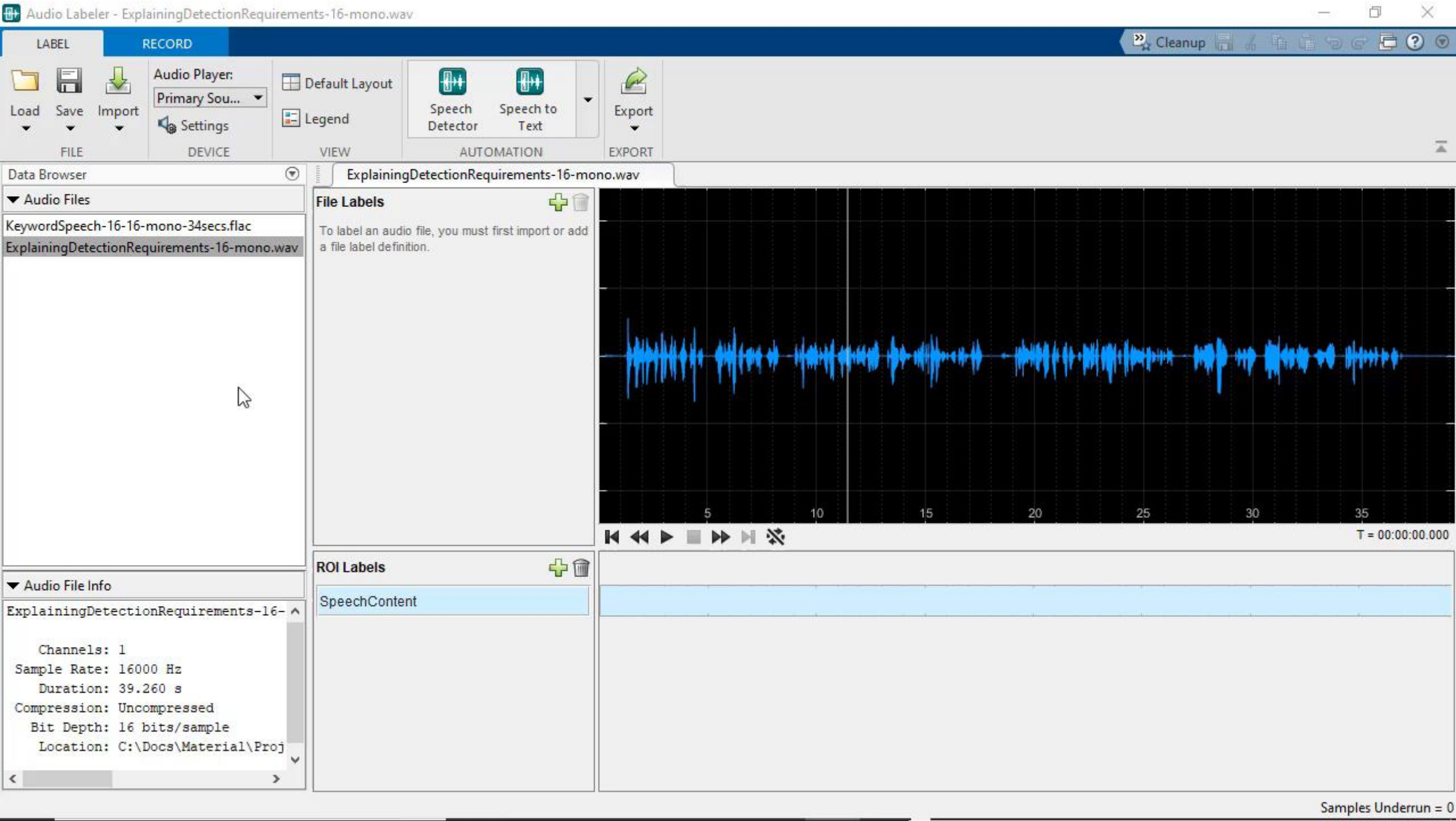

Figure 2
File Edit View Insert Tools Desktop Window Help
Validation Signal Example
1.5
1
0.5
0
-0.5
-1
-1.5
0 1 2 3 4 5 6 7
Time (s)
Speech
Keyword Mask

如何进行语音数据的标注?

使用一个经过训练的智能系统来执行标注的任务，并证明其准确性!!

例如:

- 手动标注


Audio Labeler - ExplainingDetectionRequirements-16-mono.wav
LABEL
RECORD
Cleanup
Load
Save
Import
FILE
Audio Player:
Primary Sou...
Settings
DEVICE
Default Layout
Legend
VIEW
Speech Detector
Speech to Text
AUTOMATION
Export
EXPORT
Data Browser
Audio Files
KeywordSpeech-16-16-mono-34secs.flac
ExplainingDetectionRequirements-16-mono.wav
Audio File Info
ExplainingDetectionRequirements-16-
Channels: 1
Sample Rate: 16000 Hz
Duration: 39.260 s
Compression: Uncompressed
Bit Depth: 16 bits/sample
Location: C:\Docs\Material\Proj
ExplainingDetectionRequirements-16-mono.wav
File Labels
To label an audio file, you must first import or add a file label definition.
5
10
15
20
25
30
35
T = 00:00:00.000
ROI Labels
SpeechContent
Samples Underrun = 0

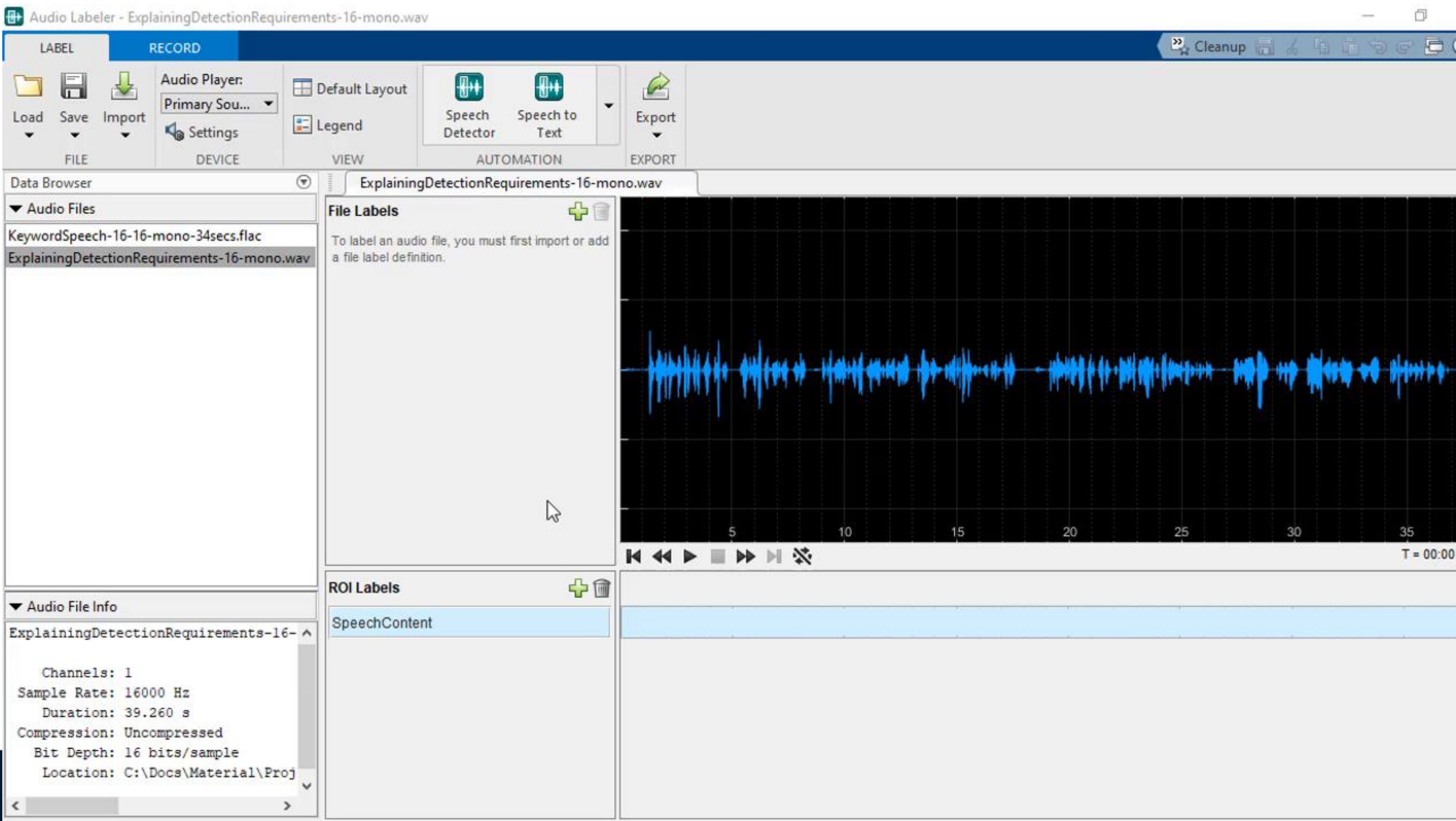
如果进行数据的标?

使用一个经过训练的智能系统来执行类似的任务，并证明其准确性!!

例如:

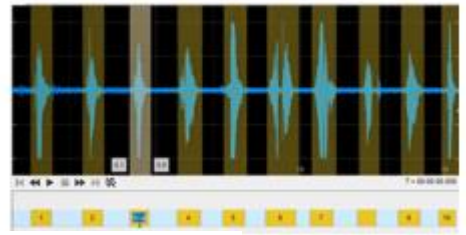
- 手动标注

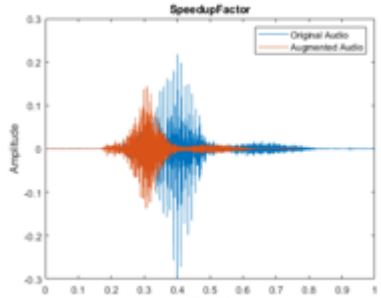
- 自动标注

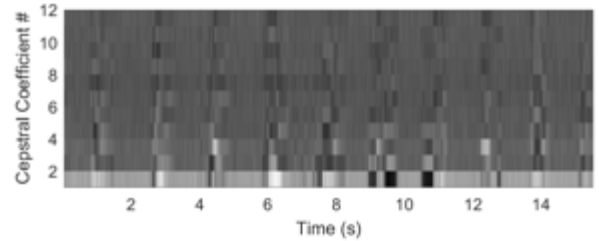

Audio Labeler - ExplainingDetectionRequirements-16-mono.wav
LABEL
RECORD
Cleanup
Load
Save
Import
FILE
Audio Player:
Primary Sou...
Settings
DEVICE
Default Layout
Legend
VIEW
Speech Detector
Speech to Text
AUTOMATION
Export
EXPORT
Data Browser
Audio Files
KeywordSpeech-16-16-mono-34secs.flac
ExplainingDetectionRequirements-16-mono.wav
Audio File Info
ExplainingDetectionRequirements-16-
Channels: 1
Sample Rate: 16000 Hz
Duration: 39.260 s
Compression: Uncompressed
Bit Depth: 16 bits/sample
Location: C:\Docs\Material\Proj
ExplainingDetectionRequirements-16-mono.wav
File Labels
To label an audio file, you must first import or add a file label definition.
5
10
15
20
25
30
35
T = 00:00
ROI Labels
SpeechContent

主要内容

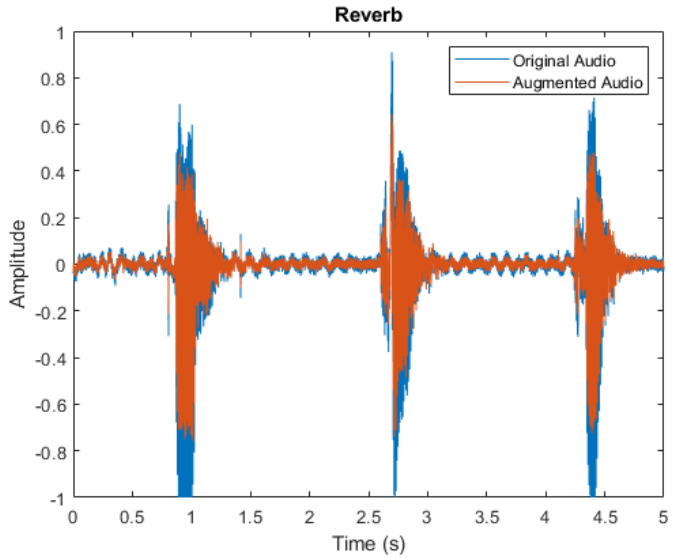
- 深度神经网络
- 数据标注
- 数据增强– 合成数据
- 输入数据转换
- 模型部署



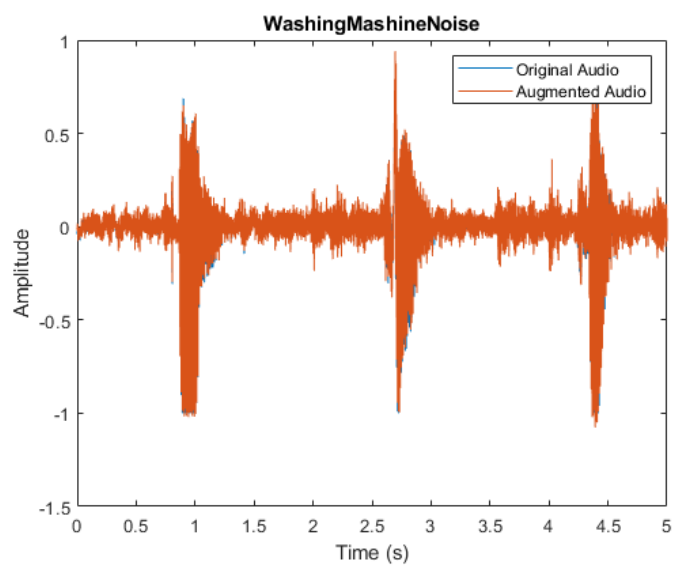




数据增强 – 有效的丰富数据集方法



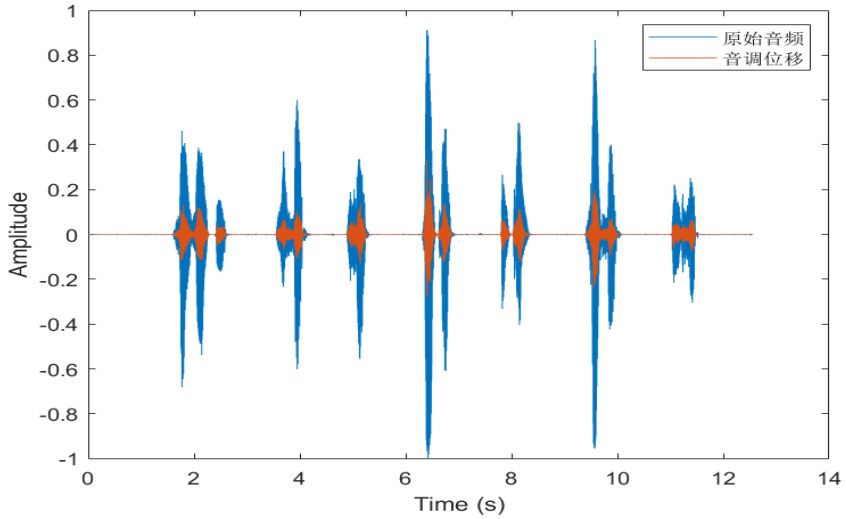
增加厨房混音



增加机器噪声

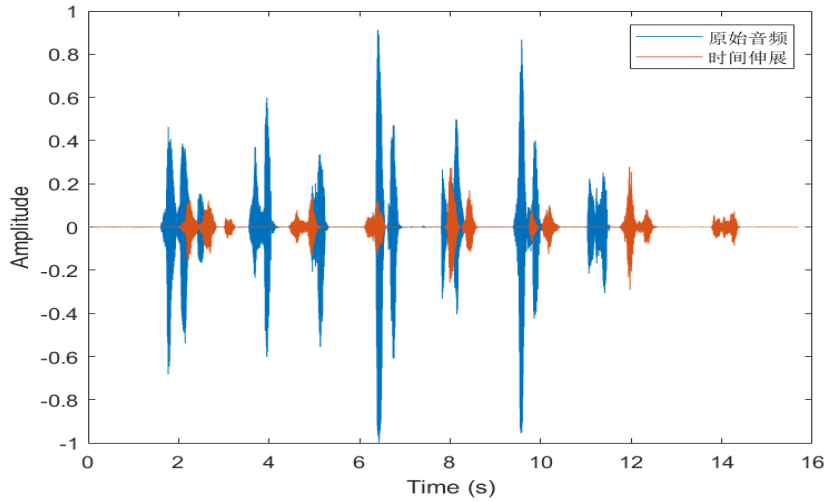
数据增强 – 有效的丰富数据集方法

```
augmenter = audioDataAugmenter( ...
    "AugmentationMode","sequential", ...
    "AugmentationParameterSource","specify", ...
    "ApplyTimeStretch",false, ...
    "ApplyPitchShift",true, ...
    "ApplyVolumeControl",false, ...
    "ApplyAddNoise",false, ...
    "ApplyTimeShift",false)
```



时间伸展(time stretch)

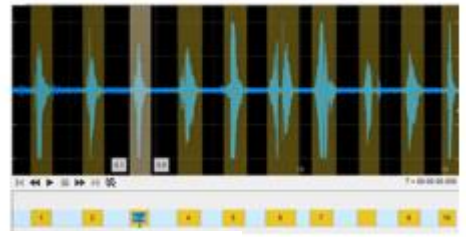
```
augmenter = audioDataAugmenter( ...
    "AugmentationMode","sequential", ...
    "AugmentationParameterSource","specify", ...
    "ApplyTimeStretch",true, ...
    "ApplyPitchShift",false, ...
    "ApplyVolumeControl",false, ...
    "ApplyAddNoise",false, ...
    "ApplyTimeShift",false)
```

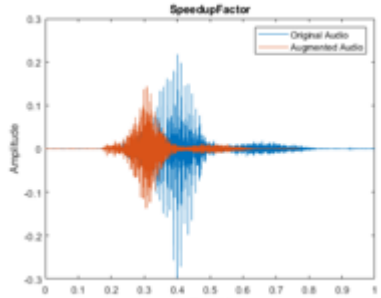


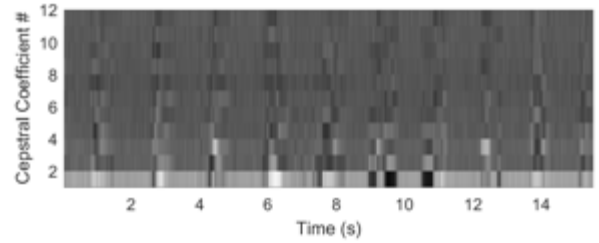
Learn more on [audioDataAugmenter]

主要内容

- 深度神经网络
- 数据标注
- 数据增强 – 合成数据
- 数据转换
- 模型部署

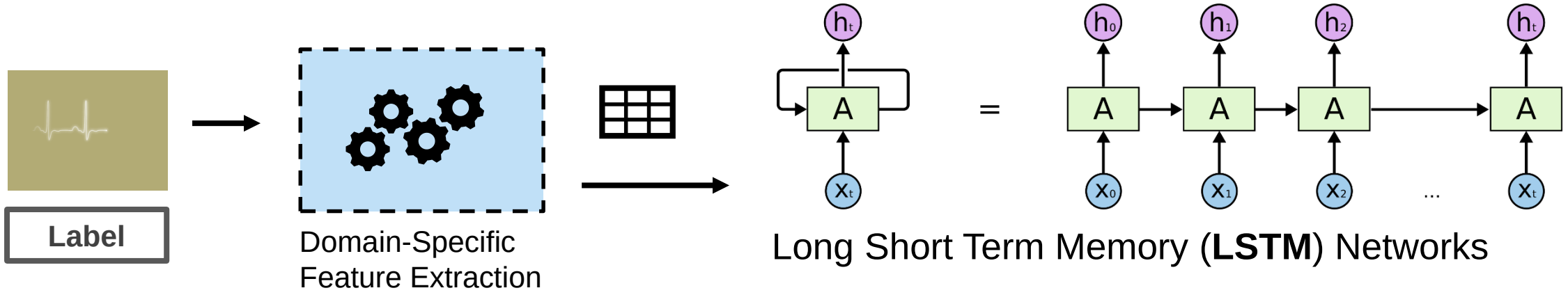






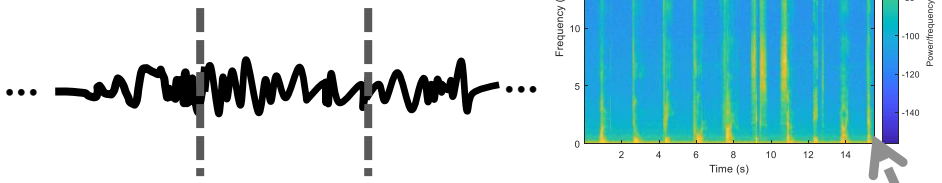
用时域信号训练深度网络通常需要提取特征

Deep learning ≠ End-to-end learning

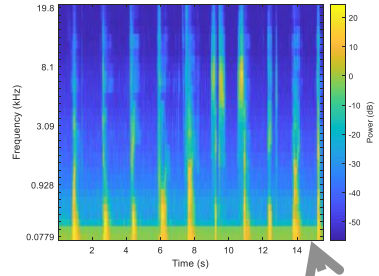


不同的应用需要不同的特征提取技术

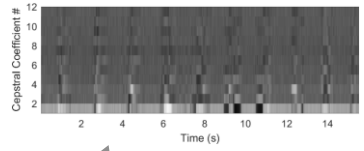
`spectrogram, stft`



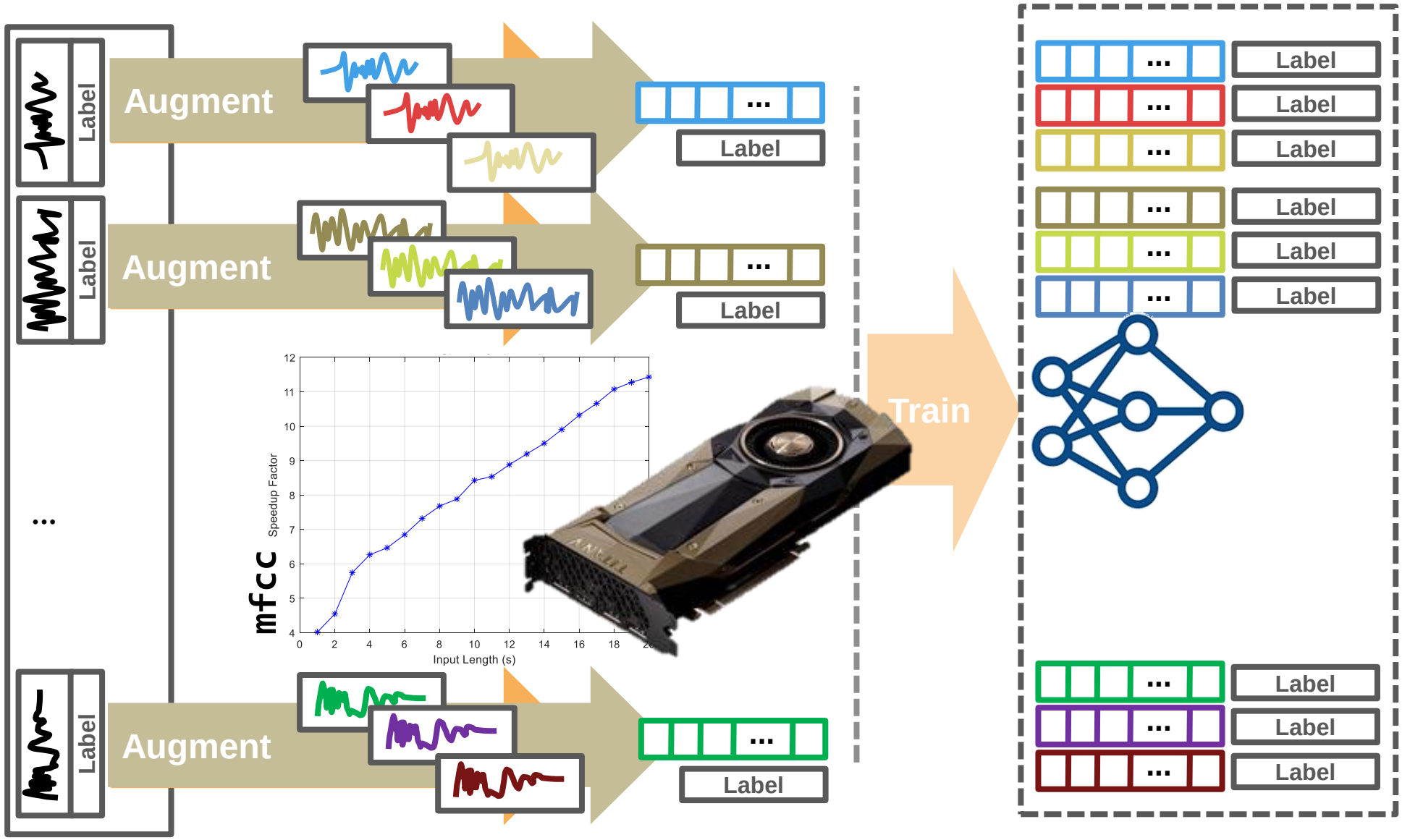
`melSpectrogram`



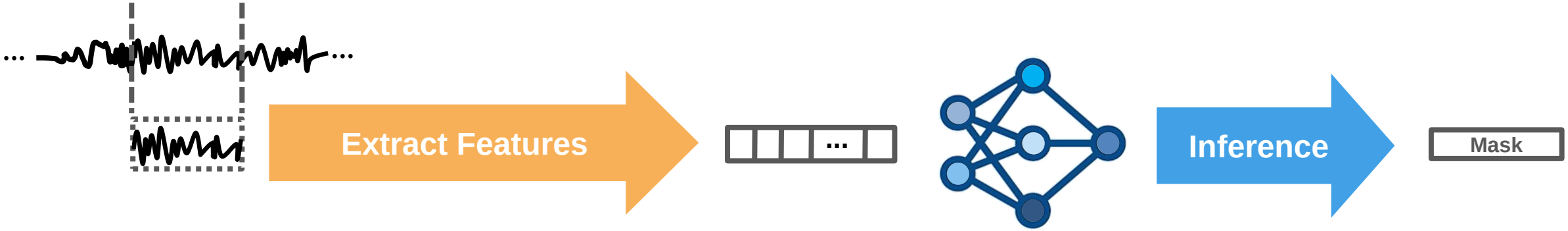
`mfcc`



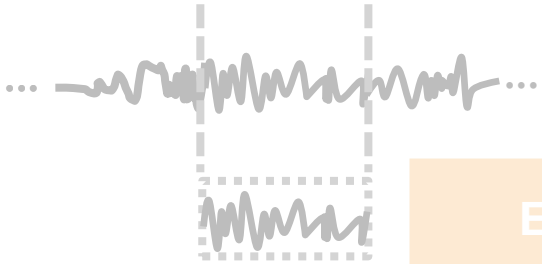
利用输入数据训练神经网络（训练）



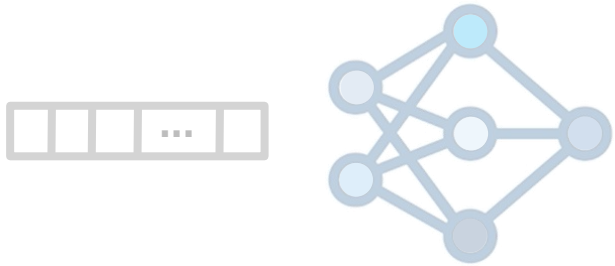
使用深度神经网络进行预测（推理）



使用深度神经网络进行预测（推理）



Extract Features



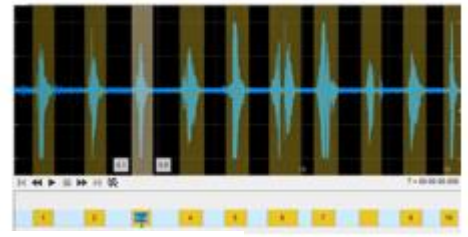
Inference

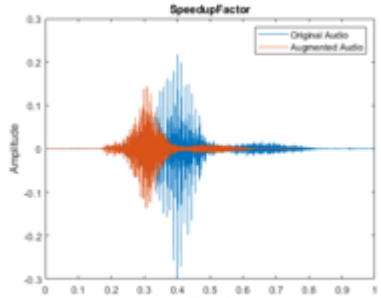
Trigger

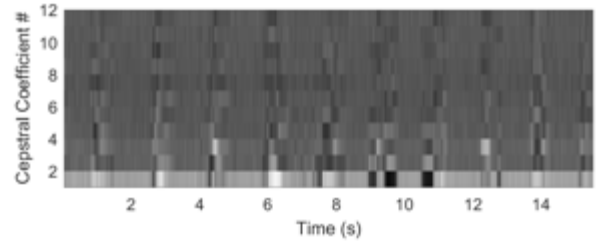

```
[...]

% Exctract MFCC from whole analysis buffer
[coeffs,delta,deltaDelta] = mfcc(buf,SampleRate,...
    'WindowLength',winLength,...
    'OverlapLength',ovlpLength);

% Concatenate and normalize features
featureMatrix = [coeffs,delta,deltaDelta];
featureMatrix = (featureMatrix - M)./S;

% Detect keyword with LSTM network (Mask around speech keyword)
featMask    = classify(net,featureMatrix.');

% Debounce and re-align detections in time domain
[timeMask, chimePosition] = debounceAnalyzeDetectionMask(featMask);

% Generate chimes for detection events
chime = generateChimeAtSample(chimePosition,...

[...]
```

主要内容

- 深度神经网络
- 数据标注
- 数据生成 – 合成数据与数据增强
- 输入数据特征转换
- 模型部署





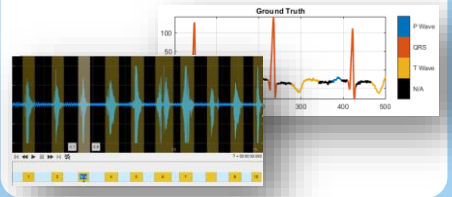


Create and Access Datasets

Data sources

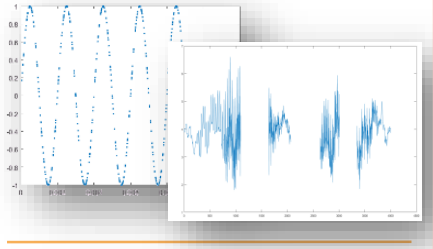
Simulation and augmentation

Data Labeling

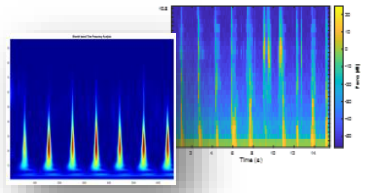


Preprocess and Transform Data

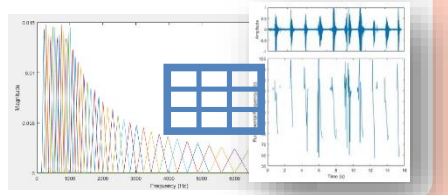
Pre-Processing



Transformation



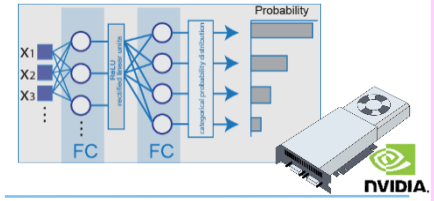
Feature extraction



Develop Predictive Models

Import Reference Models/ Design from scratch

Hardware-Accelerated Training



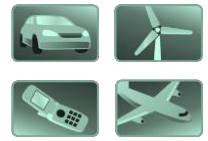
Analyze and tune hyperparameters

Accelerate and Deploy

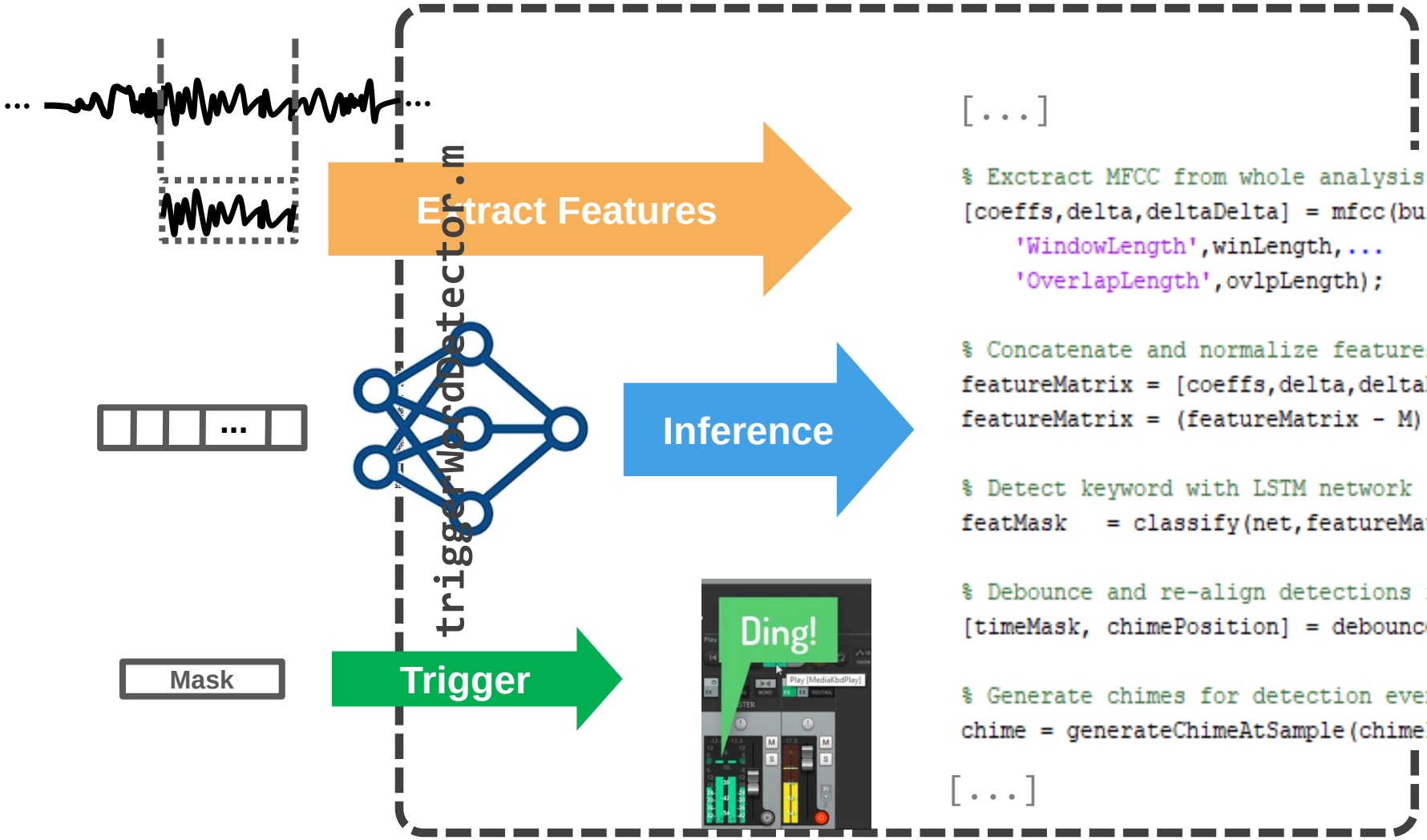
Desktop Apps

Enterprise Scale Systems

Embedded Devices and Hardware



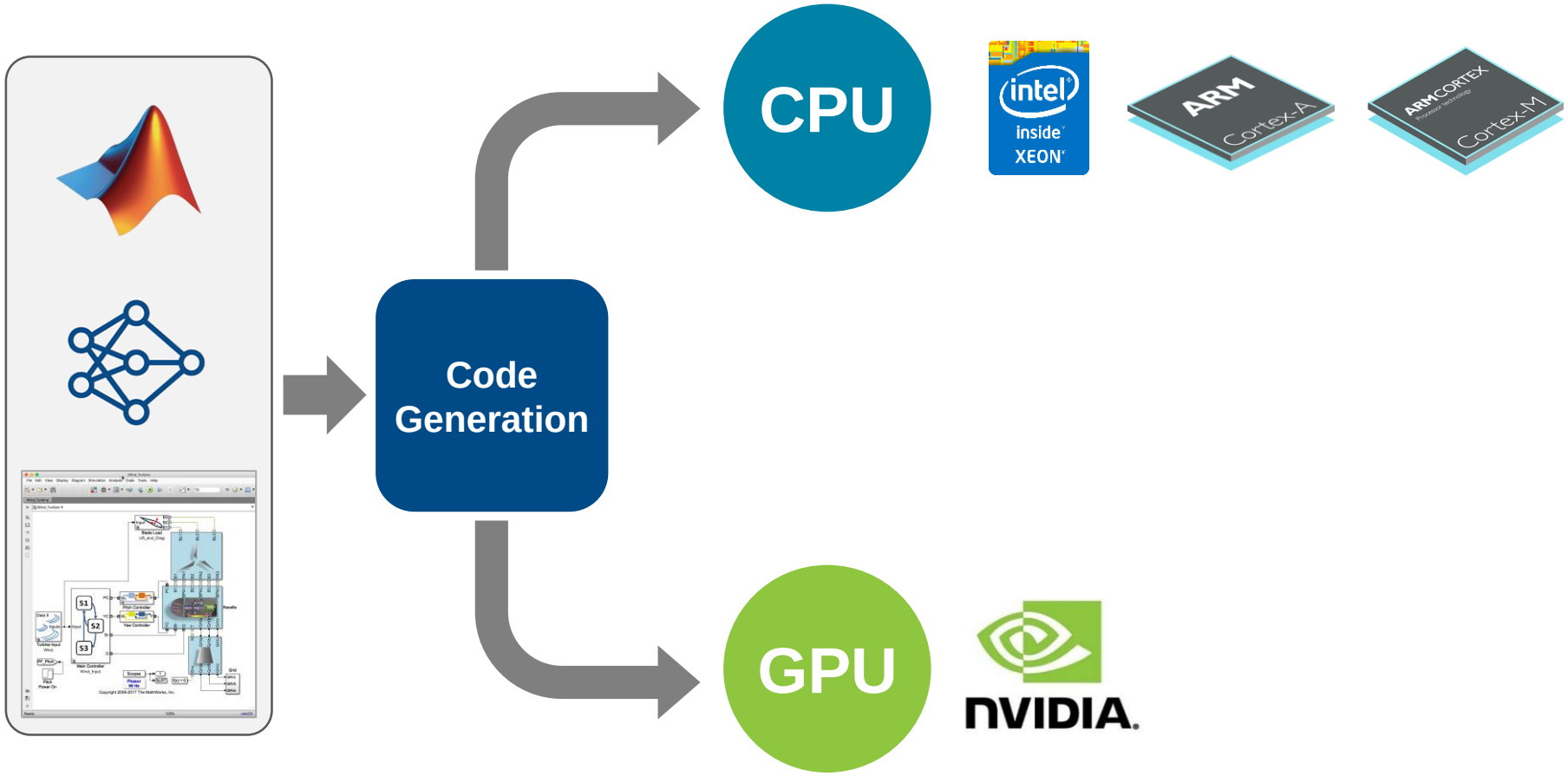
使用深度神经网络进行预测（推理）




```
[...]

% Exctract MFCC from whole analysis buffer
[coeffs,delta,deltaDelta] = mfcc(buf,SampleRate,...
    'WindowLength',winLength,...
    'OverlapLength',ovlpLength);

% Concatenate and normalize features
featureMatrix = [coeffs,delta,deltaDelta];
featureMatrix = (featureMatrix - M)./S;

% Detect keyword with LSTM network (Mask around speech keyword)
featMask    = classify(net,featureMatrix.');

% Debounce and re-align detections in time domain
[timeMask, chimePosition] = debounceAnalyzeDetectionMask(featMask);

% Generate chimes for detection events
chime = generateChimeAtSample(chimePosition,...

[...]
```

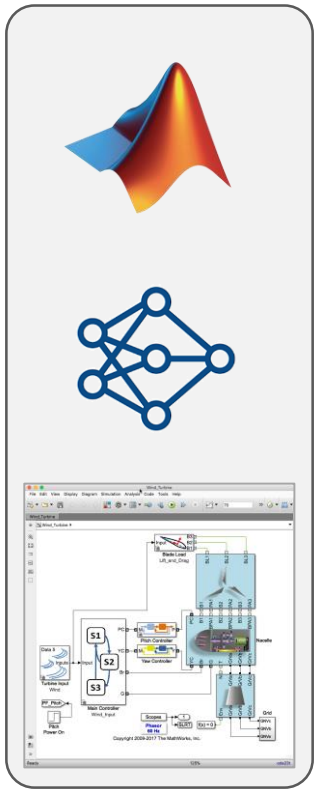
部署到多处理器上

MATLAB人工智能模型可以部署在嵌入式设备、边缘设备、企业系统、云或桌面。.



部署到多处理器上

MATLAB人工智能模型可以部署在嵌入式设备、边缘设备、企业系统、云或桌面。.



Compiler
Compiler SDK

可执行程序

第三方语言库

Web App

Web API

Windows, Linux

JAVA, C#, C++,
Python, …

Web App Server

MATLAB Production Server

语音识别应用

- **训练数据集**

  来自30个不同人员的语音数据，包含5个关键词

- **深度神经网络模型**

  双向LSTM

- **识别的语音关键词**

  北京

- **部署方式**

  桌面应用程序


Q
A