ISTAR IUL
INFORMATION SCIENCES, TECHNOLOGIES
AND ARCHITECTURE RESEARCH CENTER
CENTRO DE INVESTIGAÇÃO EM CIÊNCIAS
DA INFORMAÇÃO, TECNOLOGIAS E ARQUITETURA
INSTITUTO UNIVERSITÁRIO DE LISBOA

ASR – Automatic Speech Recognition for European Portuguese with the Kaldi Framework

Mauricio Breternitz
Miguel Sales Dias
Pedro Santos
04.JUL.2018

Outline

- Introduction
- ASR – Automatic Speech Recognition
- The Kaldi ASR Toolkit
- European Portuguese ASR Development
- Experimental Set UP, Results
- Future Work

funded by Marie Curie IRIS (ref. 610986, FP7-PEOPLE-2013-IAPP)

Digital Living Spaces

- ✓ Ambient Assisted Living
- ✓ Smart cities e smart buildings
- ✓ Mobilidade sustentável
- ✓ Realidade virtual e aumentada para edifícios
- ✓ Sustentabilidade energética
- ✓ Projeto participativo e fabricação digital
- ✓ Análise do espaço, perceção e uso
- ✓ Projeto de espaços, do património às práticas vernaculars e às gramáticas da forma

21 researchers
5 PhD Students

Software Systems Engineering

- ✓ Engenharia de software empírica
- ✓ Cibersegurança, Gestão e Proteção de ativos digitais
- ✓ Big Data e Visual Analytics
- ✓ Gestão de fluxo de trabalho científico
- ✓ Governância das tecnologias da informação
- ✓ Smart cities e smart buildings
- ✓ Smart grids
- ✓ BPM - Business Process Management

20 researchers
3 PhD Students

Information Systems

- ✓ Teorias e modelos de sistemas de informação
- ✓ Business Intelligence & Analytics
- ✓ Business Process Management
- ✓ IT Governance & Management
- ✓ E-Government
- ✓ IS aplicada a e-learning, gamification, marketing, hospitalidade e tourismo

31 researchers
11 PhD Students

Complexity and Computational Modelling

- ✓ Análise multi critério e teoria da decisão
- ✓ Geração automática de desenho/projeto
- ✓ Gramáticas da forma genéricas
- ✓ Estruturas de significado em redes de comunicação
- ✓ Análise de fenómenos sociais complexos

13 researchers
2 PhD Students

Brief BIO, Publications, Patents

PhD – Carnegie-Mellon, ECE
MSc – UNICAMP/Brazil
BSc – ITA-Brazil

Work: IBM Research, Motorola, Times N, Intel Labs, AMD Research

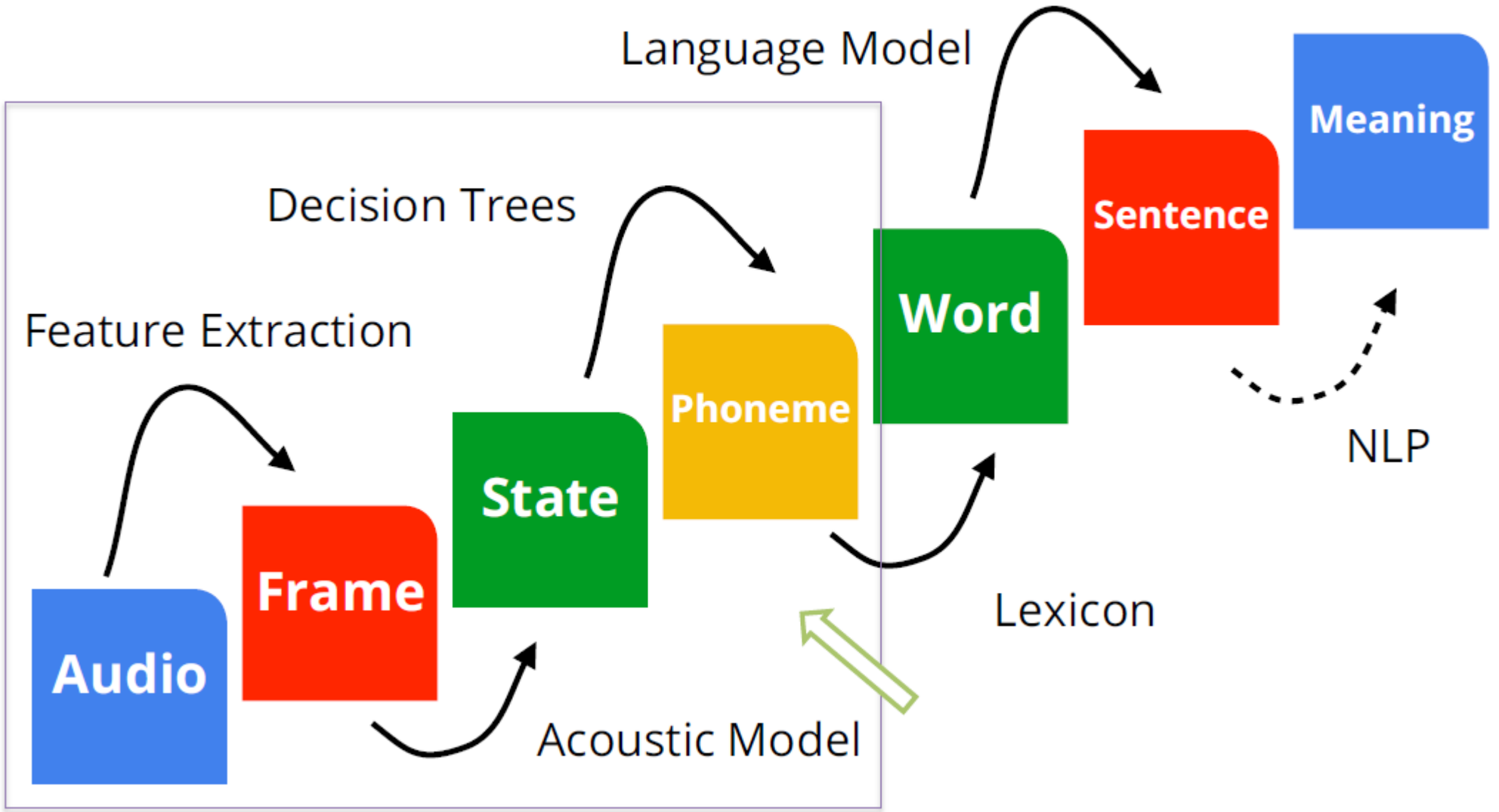
48 U.S. Patents Issued, 54 U.S. Patents Pending

Publications

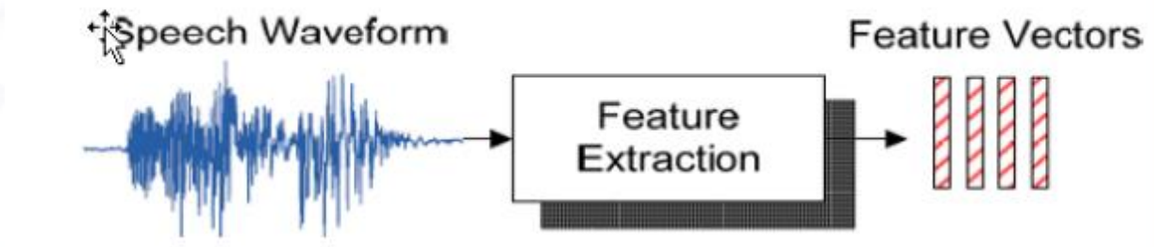
Citations 1587 H-index 24, i10-index 39
Computer Architecture, Computer Systems, Performance, Tuning
Big Data, Machine Learning

Creator /General Chair :International Workshop on Architectural/Microarchitectural Support for Binary Translation, joint with ISCA (ACM/IEEE International Symposium on Computer Architecture) and CGO.

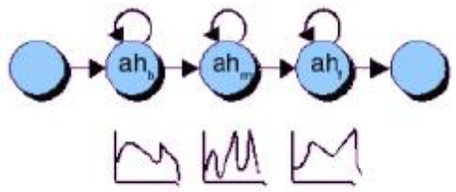
Automated Speech Recognition



* Slide from V. Vanhoucke, ICML 2013 Keynote



Phone HMM



$$\hat{W} = \text{argmax}_w P(W|O)$$

Automatic Speech Recognition

Find the most likely sentence (word sequence) $\boldsymbol{W}$, which transcribes the speech audio $\boldsymbol{A}$:

$$\widehat{\boldsymbol{W}} = \underset{\boldsymbol{W}}{\text{argmax}}\, P(\boldsymbol{W}|\boldsymbol{A}) = \underset{\boldsymbol{W}}{\text{argmax}}\, P(\boldsymbol{A}|\boldsymbol{W})P(\boldsymbol{W})$$

- Acoustic model $P(\boldsymbol{A}|\boldsymbol{W})$
- Language model $P(\boldsymbol{W})$

Training: find parameters for acoustic and language model separately

- Speech Corpus: speech waveform and human-annotated transcriptions
- Language model: with extra data (prefer daily expressions corpus for spontaneous speech)

KALDI ASR Framework

- State-of-the-art ASR framework
- Open-Source, Forum
- DNN Framework
- Components
  - processing a speech waveform,
  - language models and
  - transcriptions

EUROPEAN PORTUGUESE ASR Development:

1. Simple vocabulary (spoken digits)
   - Develop, tune scripts
2. European Portuguese(progressively larger) corpus 4 versions of 154,558 utterances total 618,232 utterances
   - 800+ hours of transcribed speech
   - Adapted ‘Fisher-English’ -> LusoFisher
3. Second training and experiment
   - combination of all available utterances
   - total of 589,540 unique utterances.

Experimental Setup

- CPU
  - Intel(R) Xeon(R) CPU E5-2660
    - 16 cores –
    - 16Gb memory –
    - 2Tbyte disk storage
- GPU: NVIDIA Corporation Titan XP (*)

  Software: CUDA 8.0
- Operating System: Ubuntu 16.04.03 LTS

(*)The authors acknowledge equipment donation by NVIDIA academic program and by Advanced Micro Devices

Experimental Corpus

CORPUS	Number of Utterances
Corpus Subset	1000
SpeechCorpus1	154558
Ensemble Dataset	588162
SpeechCorpus1 Data Augmented (4x)	618232

Corpus Input Example

Words (size 148249):

abafaneticamente 190
abafante 191
abafar 192
abafação 193
abafeira 194
abafo 195
abafura 196
abafável 197
abaganhado 198
abaganhar 199
…

Lexicon:

<UNK> SPN
a a
a aex
í i
à a
à-direita a d i dx aex j t aex
à-distância a d i sh t aexn s j aex
à-esquerda a sh k e dx d aex
àgua a g w aex
….

Text:

```
SA001EP2  maria inês pedroso
SA001EP3  rosmaninho
SA001EQ1  não
SA001ET1  às sete
SA001ET2  às oito
SA001F01  olá inês é a célia que está a falar estou te a mandar esta mensagem a ver
comigo a uma a uma festa logo à noite que é da da patrícia  é os anos da patrícia e
e vai lá estar toda a gente a ver se nós nos encontramos com todos
SA001F02  olá boa tarde o meu nome é célia aleixo e eu gostaria de saber qual é que
da partida do para do voo para para as canárias e se há neste mês se não há ou se e
viagem se leva muito muito tempo se não pronto
```

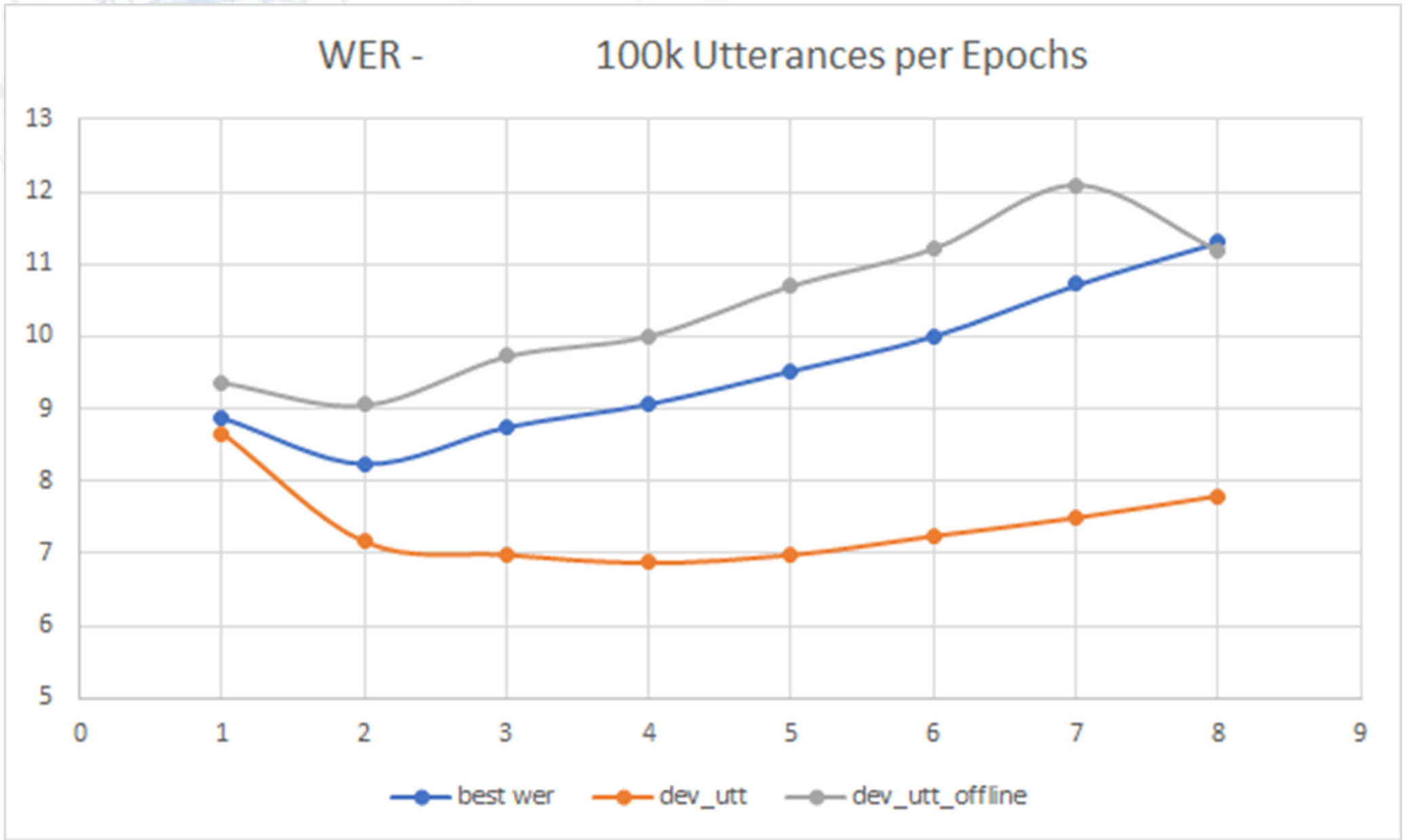
Training – Computational Requirements

CORPUS	Total Elapsed Processing Time(hours)
SpeechCorpus1 Subset	12.5
SpeechCorpus1	194
SpeechCorpus1 Data Augmented (4x)	620

Trained Model Size

CORPUS	Trained Model Size (Mbytes)
SpeechCorpus1 Subset	47
SpeechCorpus1	48
Ensemble Dataset	97
SpeechCorpus1 Data Augmented (4x)	91

Training vs OverFitting



Results- WER

CORPUS	DNN Best WER	Traditional Best WER
SpeechCorpus1 Subset	10.46	12.75
SpeechCorpus1	11.77	15.61
Ensemble Dataset	10.12	13.17
SpeechCorpus1 Data Augmented (4x)	15.99	18.23

<<Demo>>

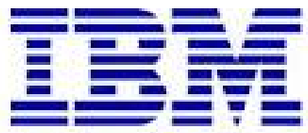
- Kaldi -> Library

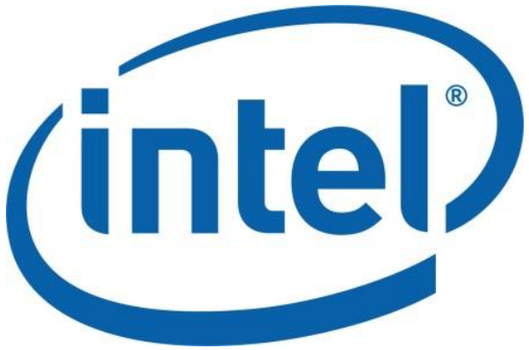
Conclusion & Future Work

- End-to-End ASR
- Research topics:
  - Full-stack: HW, SW, Application, System
  - Cross-layer optimization
  - Societal Applications

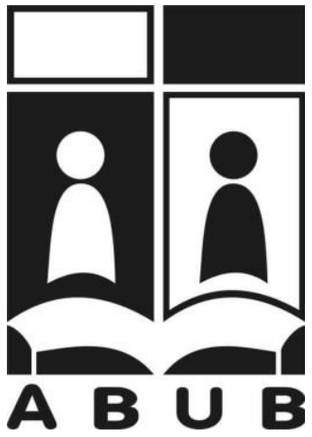
education

work





personal



ISCTE IUL
Instituto Universitário de Lisboa

Backup

End-To-End ASR

End-to-End Speech Recognition From the Raw Waveform

Neil Zeghidour[1,2], *Nicolas Usunier*[1], *Gabriel Synnaeve*[1], *Ronan Collobert*[1], *Emmanuel Dupoux*[2]

[1] Facebook A.I. Research, Paris, France; New York & Menlo Park, USA
[2] CoML, ENS/CNRS/EHESS/INRIA/PSL Research University, Paris, France
{neilz, usunier, gab, locronan}@fb.com, emmanuel.dupoux@gmail.com

STATE-OF-THE-ART SPEECH RECOGNITION WITH SEQUENCE-TO-SEQUENCE MODELS

Chung-Cheng Chiu, Tara N. Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J. Weiss, Kanishka Rao, Katya Gonina, Navdeep Jaitly, Bo Li, Jan Chorowski, Michiel Bacchiani

Google, USA
{chungchengc,tsainath,yonghui,prabhavalkar,drpng,zhifengc,anjuli
ronw,kanishkarao,kgonina,ndjaitly,boboli,chorowski,michiel}@google.com

ABSTRACT

Attention-based encoder-decoder architectures such as Listen, Attend, and Spell (LAS), subsume the acoustic, pronunciation and language model components of a traditional automatic speech recognition

on a large vocabulary continuous speech recognition (LVCSR) task. The goal of this paper is to explore various structure and optimization improvements to allow sequence-to-sequence models to significantly outperform a conventional ASR system on a voice search task.

Site ISTAR

http://istar.iscte-iul.pt/

Ciência-IUL ISTAR

https://ciencia.iscte-iul.pt/centres/istar-iul

Key Machine Learning Frameworks

CAFFE
Torch
Tensorflow
Theano
Keras

ML Services

Google ML API

Prediction API https://cloud.google.com/prediction/docs/

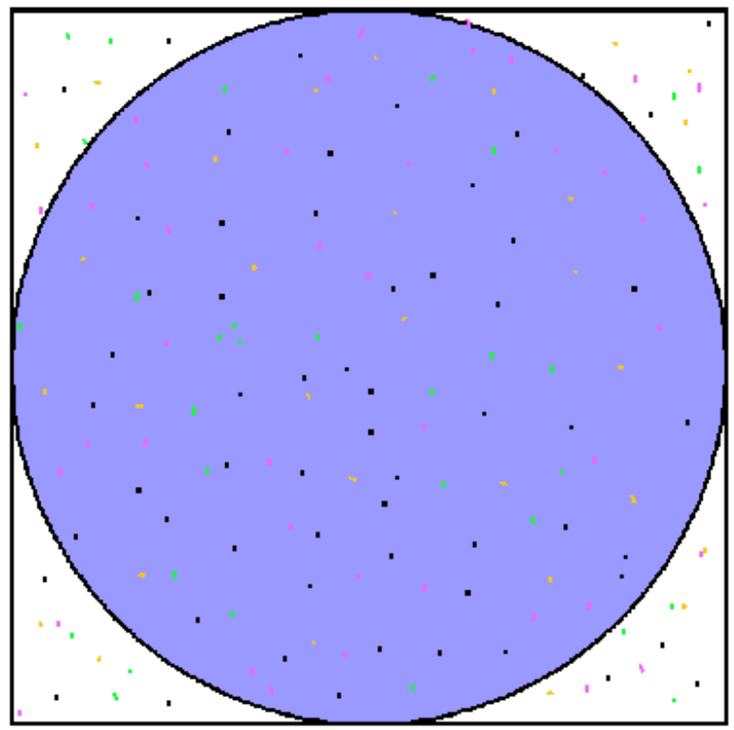
MICROSOFT COGNITIVE SERVICES

APIs https://www.microsoft.com/cognitive-services/en-us/apis
Vision, Speech, Language, Knowledge, Search

MapReduce in 30 seconds: Estimating PI

1- pick random points in unit rectangle

2- count fraction inside circle



Area: π/ 4

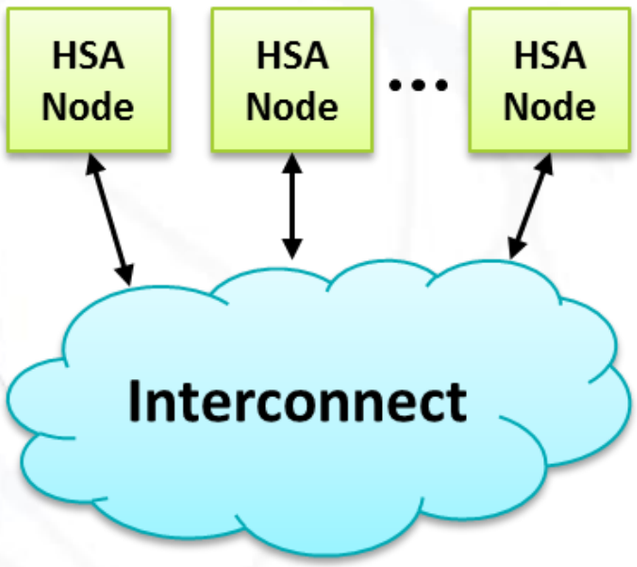
Map-Reduce:

Map: random point inside? Issue k=1, v=1 else k=0,v=1

Reduce: count 0 keys and count 1 keys

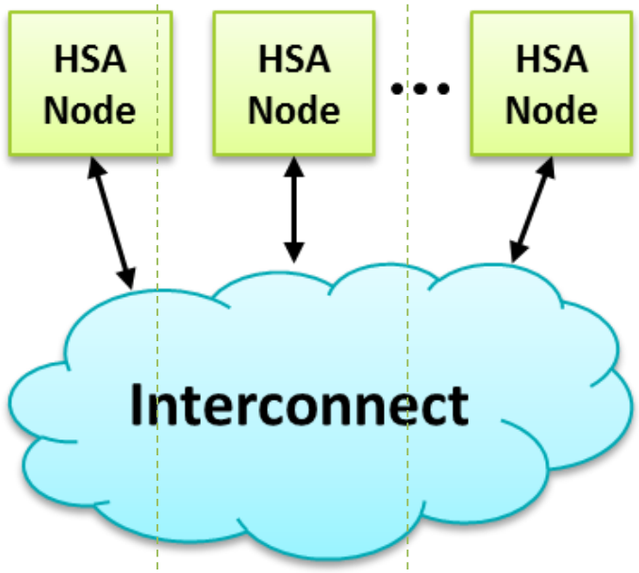
Programmer: writes { map, reduce } methods, system does rest

Nested Processing Approach



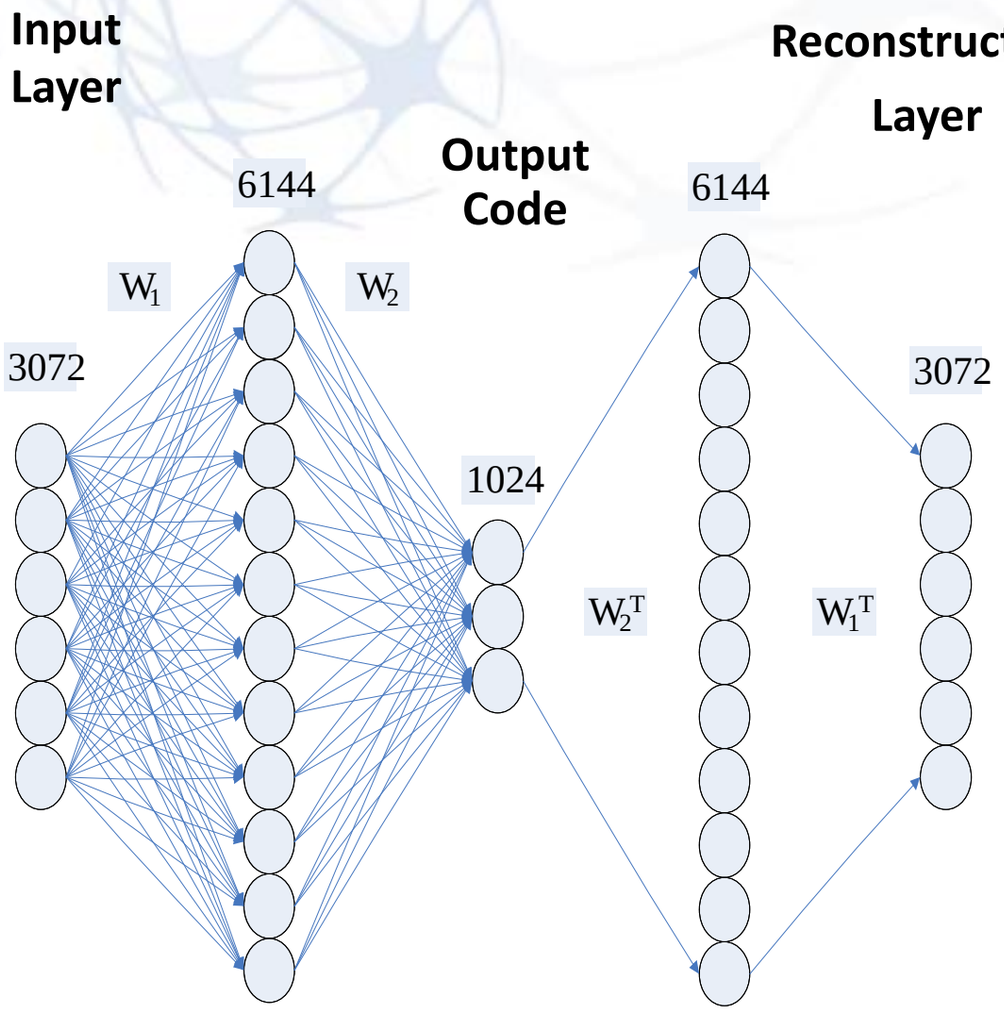
PARTITION → **WORK LOCALLY** → **COMBINE RESULTS**

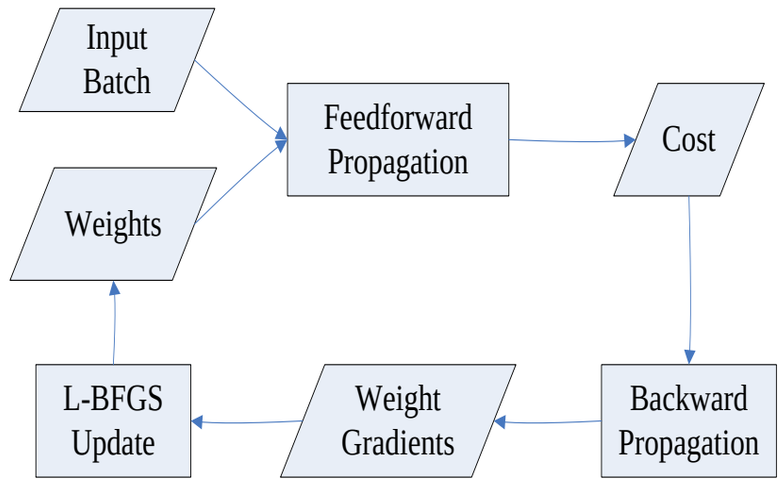
nest: CPU+GPU



DNN Autoencoder

- Autoencoder + L-BFGS training (used by Google)
  - Encode the input and then reconstruct
  - A mix of CPU compute with GPU compute
  - Heavy CPU and GPU data interaction- good match for HSA





L-BFGS Training Algorithm

End to End Speech Recognition

Feature extraction stage

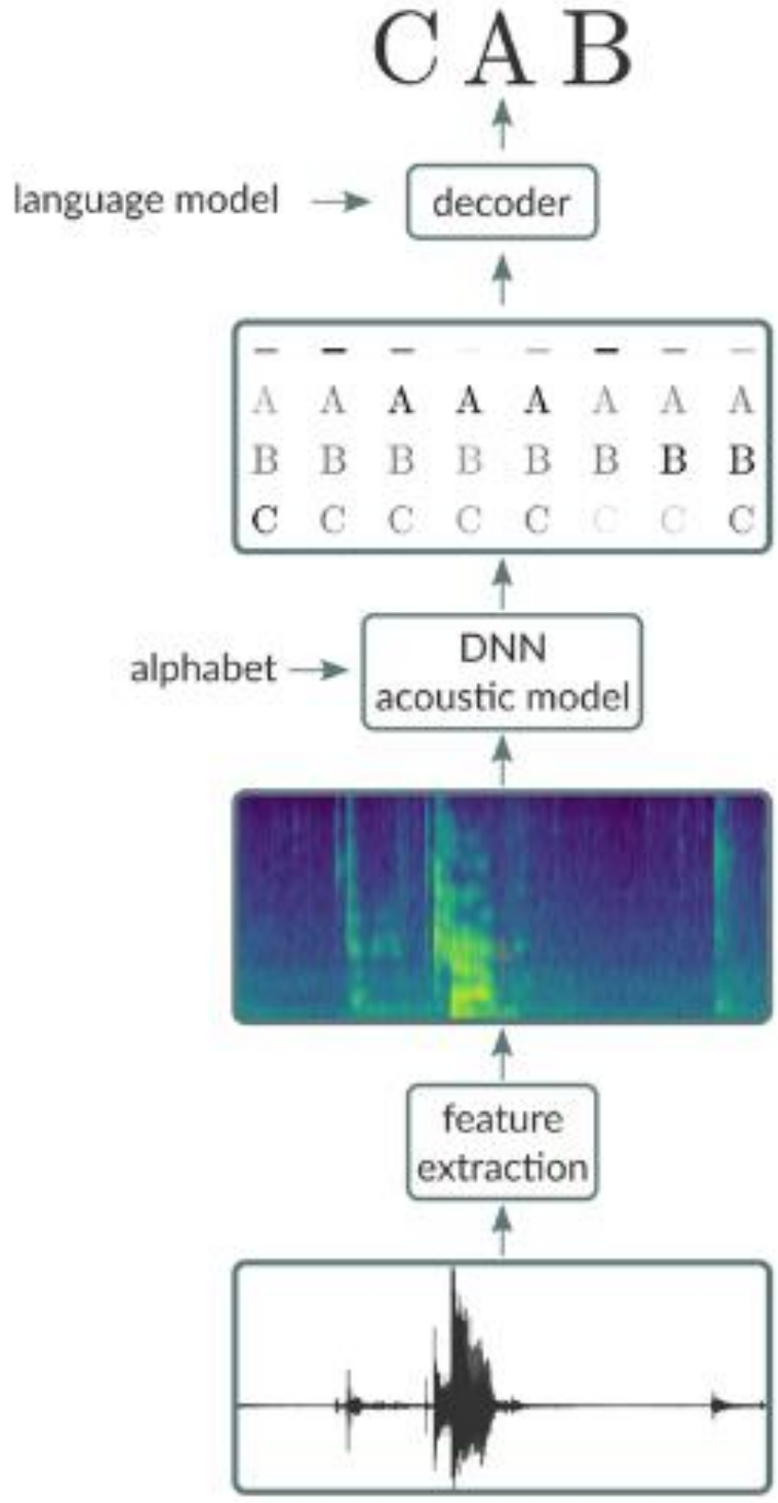
takes raw audio signals (from a wav file) and generates a sequence of feature vectors, with one feature vector for a given frame of audio input.

Acoustic model stage

takes sequences of feature vectors and generates probabilities of either character or phoneme sequences conditioned on the feature vector input.

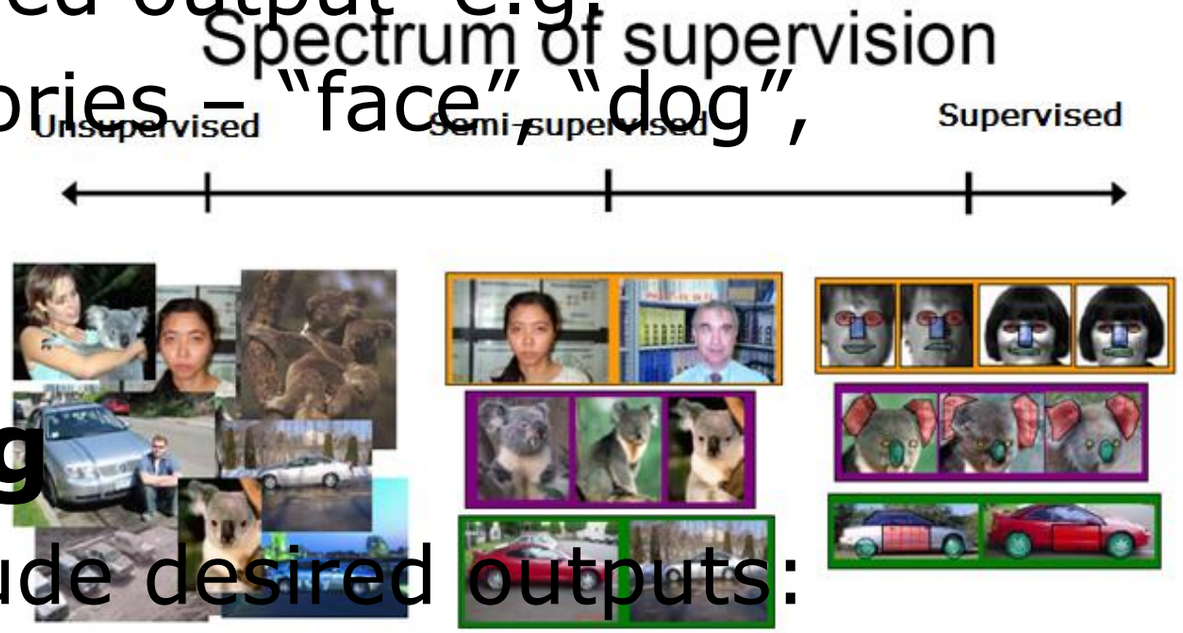
Decoder stage

takes two inputs – the acoustic model's outputs as well as a language model – and searches for the most likely transcript given the sequences generated by the acoustic model constrained by the linguistic rules encoded in the language model.



TYPES OF MACHINE LEARNING

- **Supervised (inductive) learning**
  – Training data includes desired output e.g. images are labeled with categories – “face”, “dog”, “car” …



- **Unsupervised learning**
  – Training data does not include desired outputs: the system autonomously sorts similar images and classifies data

- **Semi-supervised learning**
  – Training data includes a few desired outputs: the system **generalizes** from labeled examples

MACHINE LEARNING: APPLICATION FRAMEWORK across MULTIPLE REGIONAL SCOPES

data travel distance

Point of Data Capture	Local/Regional Data Center	Scale-Out Data Center
		
Compute reaction nearby the place data is captured, using locally stored knowledge Examples - Car warning system using front camera image; Factory Sensor Input	Compute reaction using aggregate data from a few, logically neighboring compute sites Example: Traffic management and scheduling of a city-wide fleet	Compute reaction using knowledge from a worldwide and long term memory Example: Netflix movie recommendation engine

Exascale Technology Benefits

- "Big data is what happened when the cost of storing information became less than the cost of ***making the decision*** to throw it away."
  - *George Dyson*

- Big Data and Analytics
  - Machine Learning at Exascale
  - Commercial Applications

- High Performance Computing Systems
  - Memory
  - Storage
  - Communication

The National Strategic Computing Initiative* and DOE's role

Through Executive Order 13702 he signed July 29, 2015, President Obama established the National Strategic Computing Initiative (NSCI) to maximize the benefits of HPC for US economic competitiveness and scientific discovery.

DOE is a lead agency within NSCI with the responsibility that the DOE Office of Science and DOE National Nuclear Security Administration will execute a joint program focused on advanced simulation through a **capable** exascale computing program emphasizing sustained performance on relevant applications.

* https://www.whitehouse.gov/the-press-office/2015/07/29/executive-order-creating-national-strategic-computing-initiative

ACADEMIC COLLABORATIONS

- DIVIDEND - Distributed Heterogeneous Vertically IntegrateD ENergy Efficient Data centres
  - Marie Curie CHIST-ERA: Edinburgh,Belfast, Lancaster, EPFL, Timisoara, AMD, INRIA
  - Euros 2M
  - http://www.chistera.eu/projects/dividend
- EUPHONIC – ITN 2018 Submission
- Pesquisador Visitante Especial - UNICAMP
- Invited Lecturer:

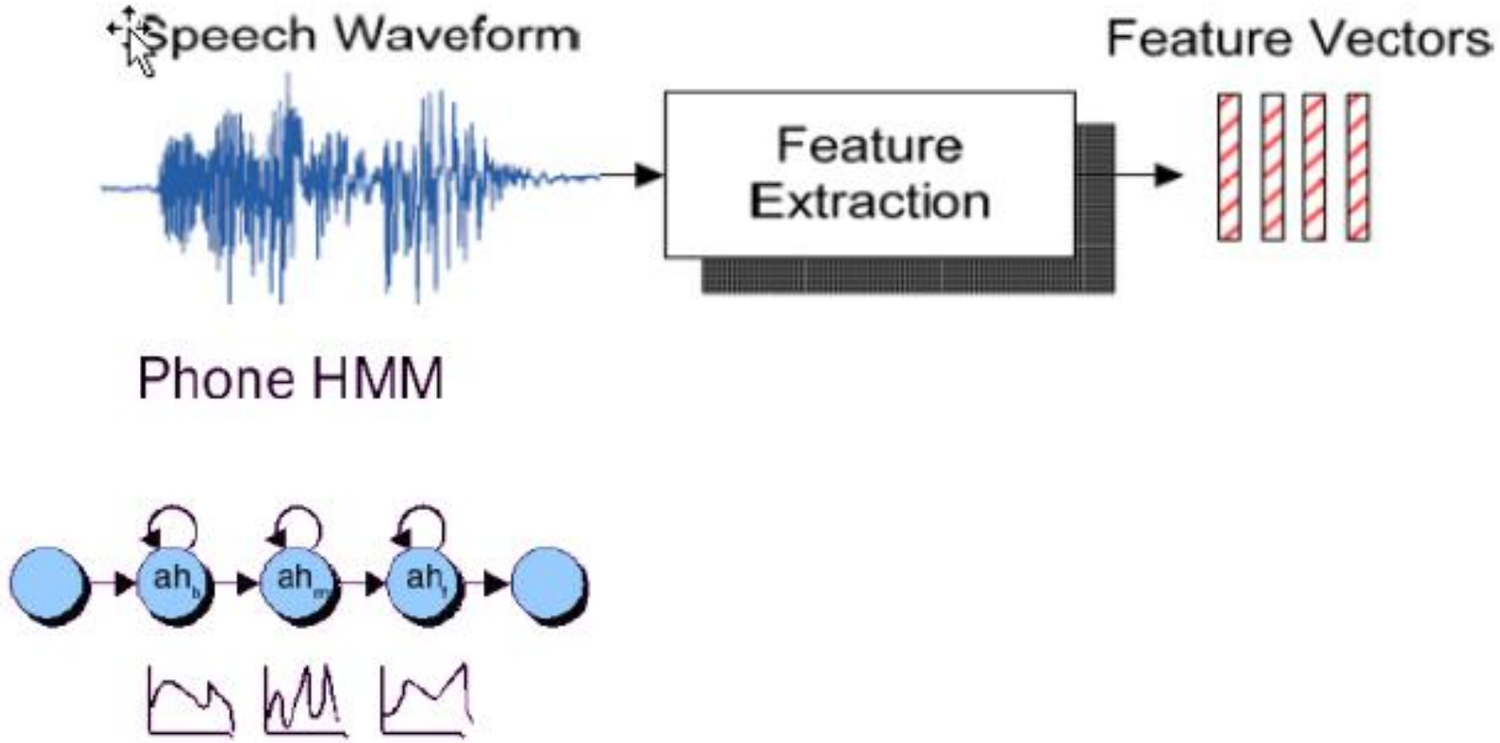
  U.Texas/Austin, Rice University, CMU-Silicon Valley

ASR System Overview

Courtesy of Bin Ma (2015)

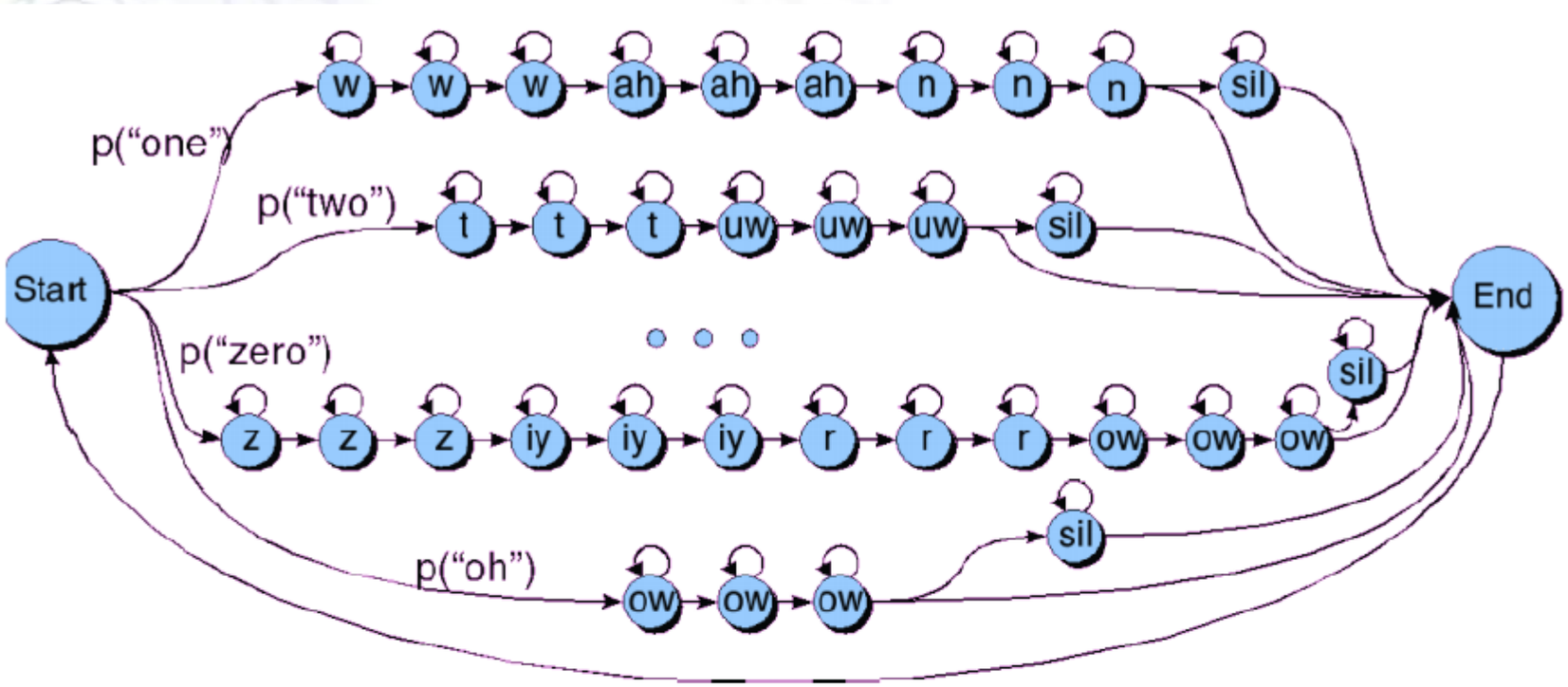
ASR In Brief

Lexicon



$$\hat{W} = \text{argmax}_w P(W|O)$$

ASR In Brief



- P(W|O) = P(O|W)P(W): P(O|W) probability of a feature sequence given a word sequence, called acoustic model (AM), P(W) word language model (LM).
- Each word is decomposed into phonemes according to a lexicon, and each phone is modeled by a 3-state left-right Hidden Markov Model.
- Conventionally, the HMM state emission probability P(o|s) is modeled by Gaussian Mixture Models (GMMs).

- LM is usually a N-gram.

ITA

UNICAMP

Carnegie Mellon University

CPqD

IBM

MOTOROLA

intel

AMD

ISCTE IUL
Instituto Universitário de Lisboa

ESPORTE CLUBE XV DE NOVEMBRO
XV
PIRACICABA
1913

ABUB

CBF
BRASIL

BSF
bible study fellowship

A casa DA cidade