A brief introduction to speech processing (classification) topics

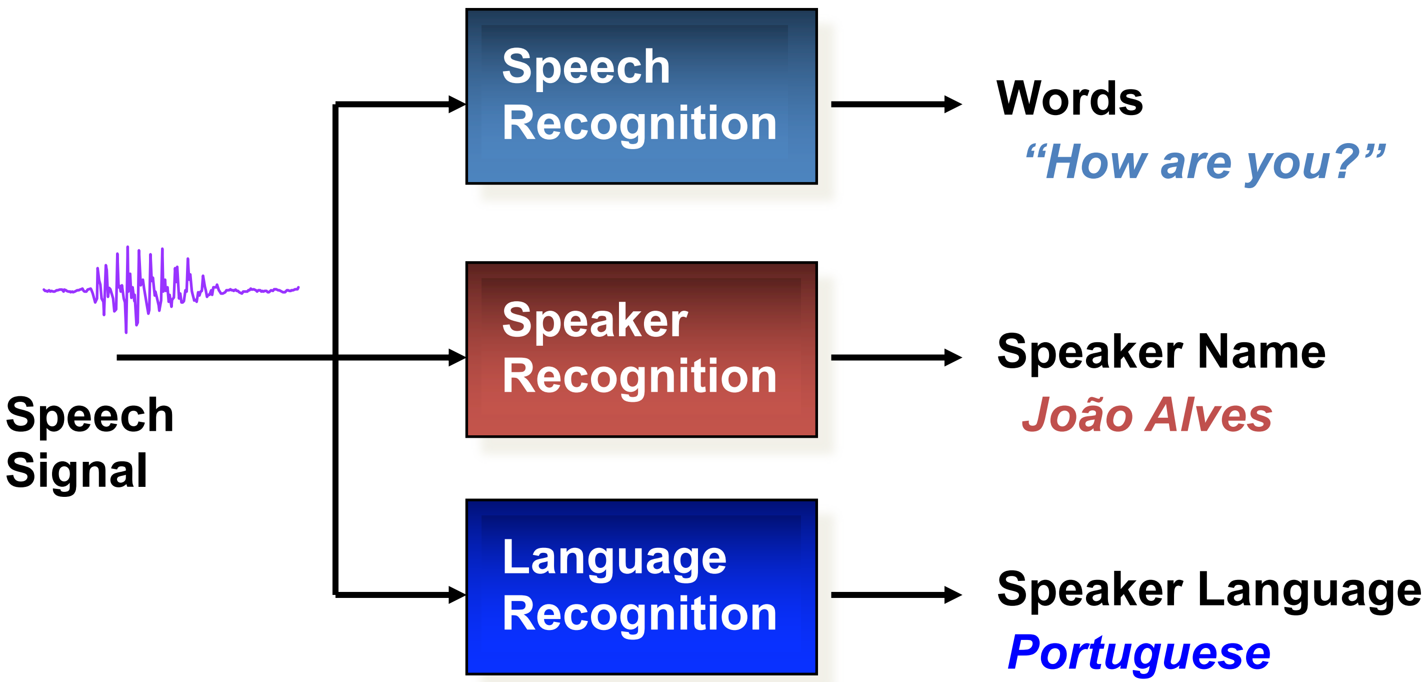
Alberto Abad
DEI – Instituto Superior Técnico – ULisboa
L²F Spoken Language Systems Lab – INESC-ID

alberto.abad@tecnico.ulisboa.pt
https://www.l2f.inesc-id.pt/w/Alberto_Abad_Gareta

Research Topics, PDEIC December 5, 2019

Introduction

Human Language Technologies



Speech processing: Speech coding, Speech enhancement, Audio segmentation, Text-to-speech synthesis, Automatic speech recognition, Speaker and language identification; Other speech pattern classification tasks

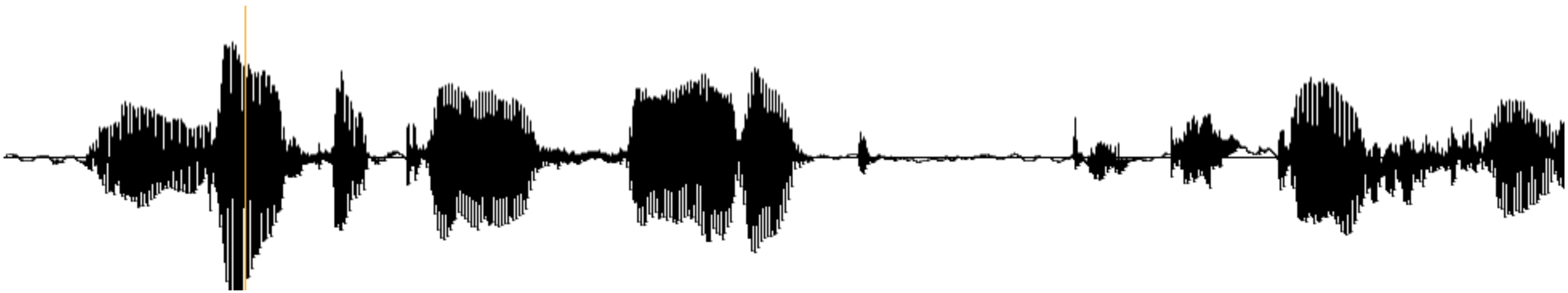
Text processing: Morphological analysis, Syntactic analysis, Semantic analysis, Discourse analysis, Named entity extraction, NL Generation, Information retrieval, Summarization, Question answering, Machine translation, Text analytics, Recommendation

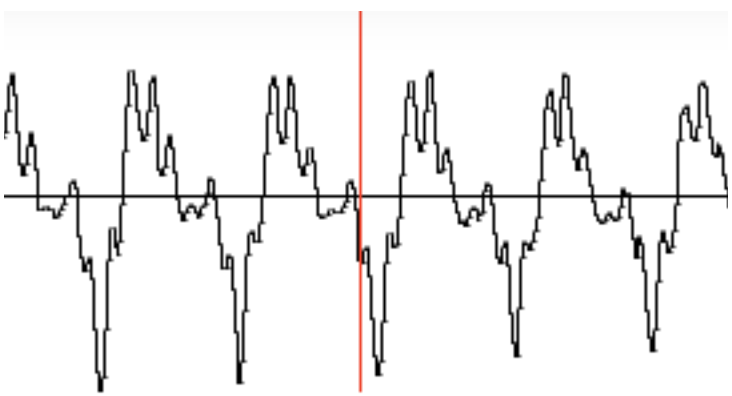
Outline

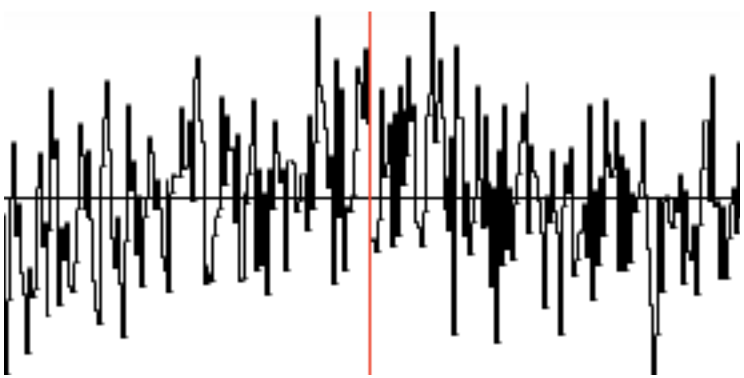
- Introduction to speech processing
- **Speech pattern classification**
- Selected research topics
- Two recent research works:
  - Domain adaptation for low-resource ASR
  - Native language (L1) identification

Speech Pattern Classification

Speech signal in the time domain

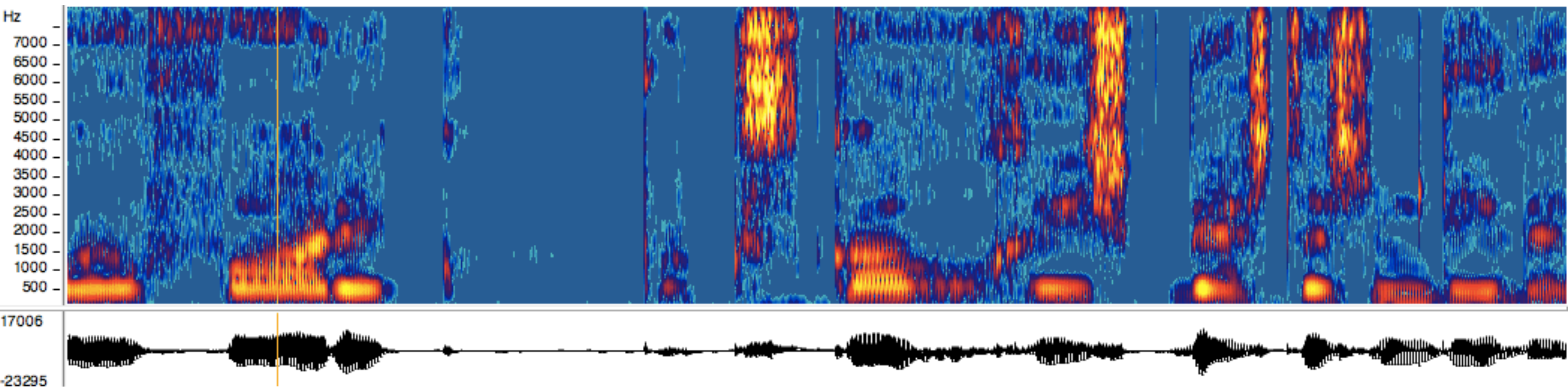


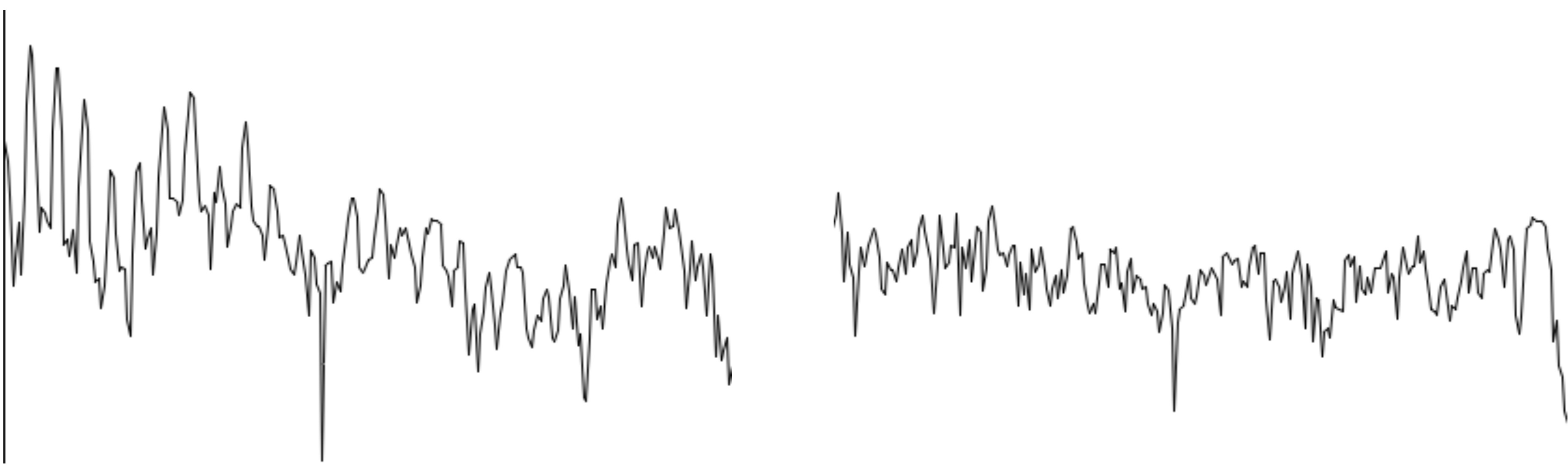




Speech Pattern Classification

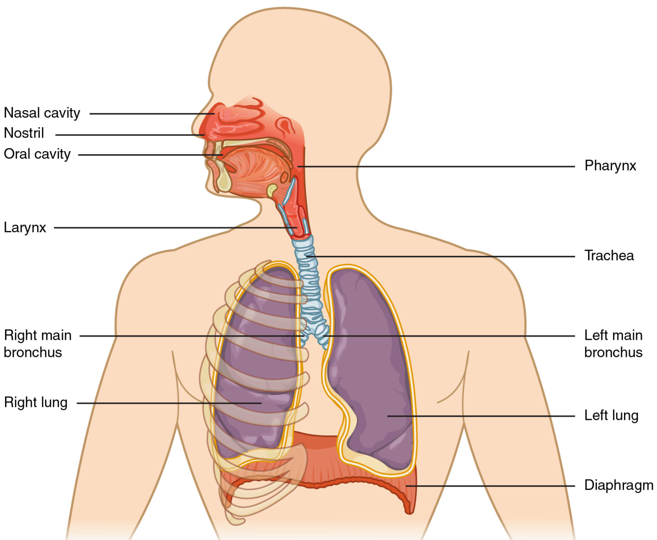
Speech signal time / frequency representation

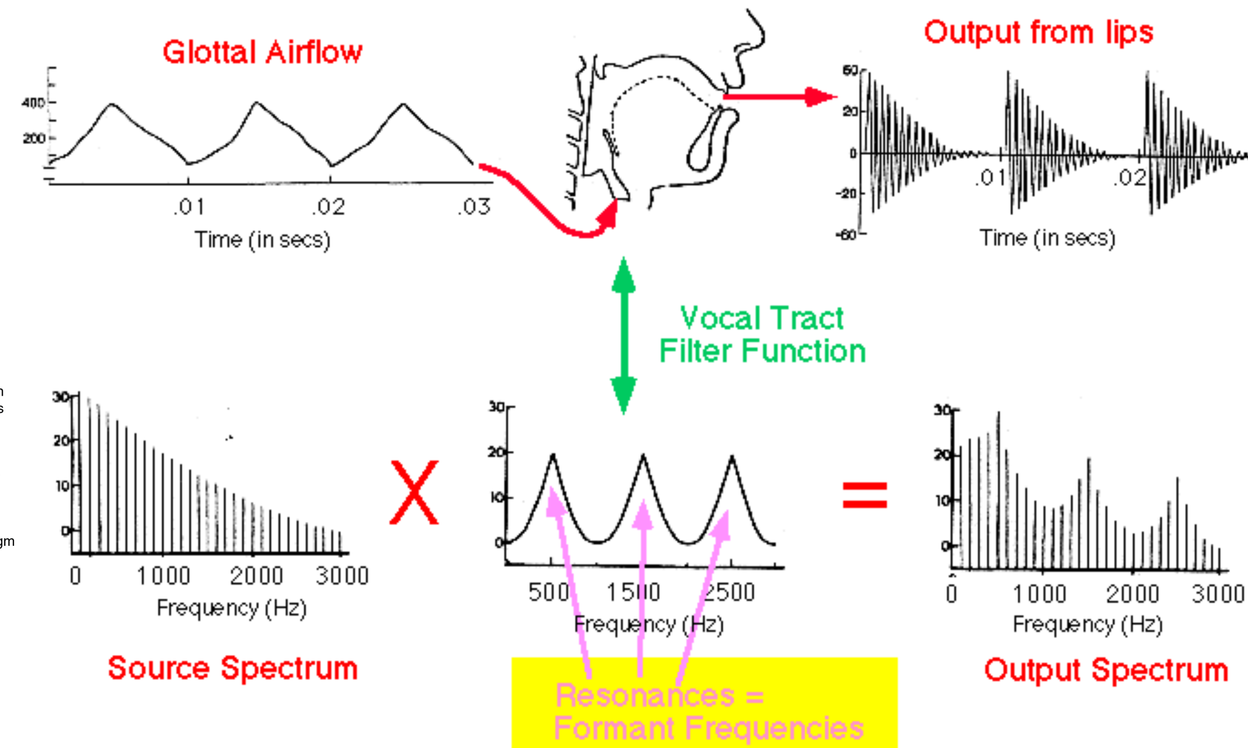




Speech Pattern Classification

Speech signal: Physiology & Source/filter model

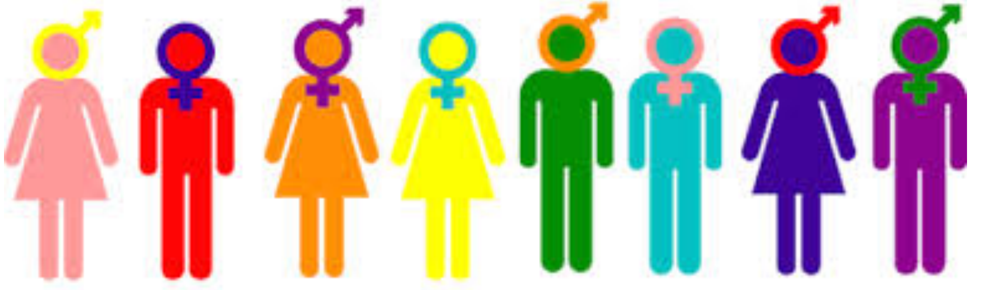




Speech Pattern Classification

Speech as a carrier of information

- Speech carries a lot of information:
  - Of course information related to the message (LINGUISTIC?)...
  - ... but also, speaker traits (NON-LINGUISTIC/PARA-LINGUISTIC?):
    - Gender; Age; Language/accent; ID; Personality; Education; Intoxication; Sleepiness; Friendliness; Mood; Physical Stress; Cognitive Load; Emotion; Pathologies?



- If "Speech" is considered in a wider sense ("Audio") then more information is present:
  - Number of speakers; speakers role; speaker position; audio events; acoustic Scenes;

Speech Pattern Classification

Speech as a carrier of information

Wouldn’t it be dreamy if only we could extract all these information automatically!?!? Then, we could imitate human behavior (IA); or even augment capabilities (data mining); or...

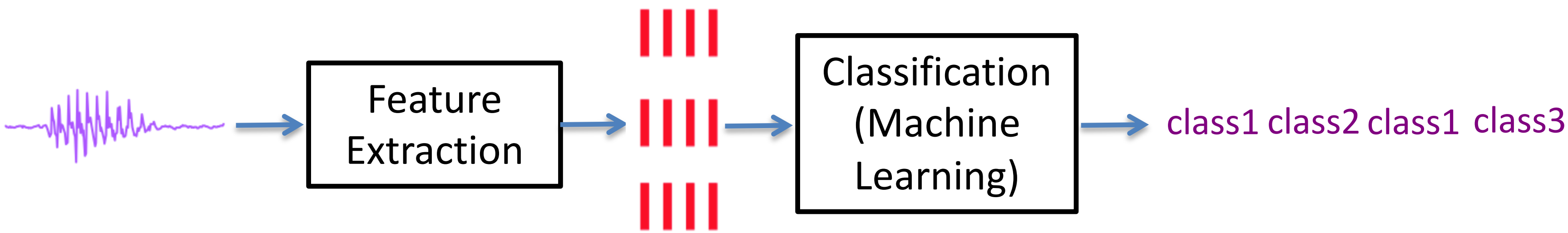
YES, we can!!! (more or less)

→ **SPEECH PATTERN CLASSIFICATION**

Speech Pattern Classification

Objectives

- The objective of ***speech pattern classification*** is to convert a speech input sequence into a sequence of class labels:



- The common blocks of any speech pattern classification task are the front-end/feature extraction and the back-end/classification:
  - The classifier module is "learnt" using data during the training phase and used to classify new unseen data during test
- Some examples are:
  - Automatic speech recognition; speech segmentation; speaker recognition; language recognition; speaker diarization; automatic document indexing; paralinguistic speaker trait recognition

Speech Pattern Classification

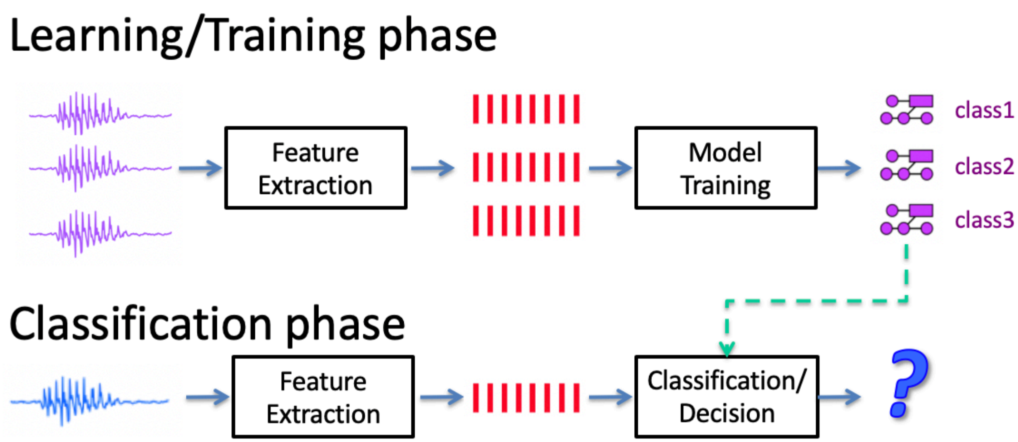
Challenges

- Speech/audio variability → Samples belonging to the same "class" take extremely different forms due to:
  - Source variation: speaker, gender, accent, state, volume, etc.
  - Channel variation: mikes, acoustic environment, noise, reverberation, etc.
  - Other: Intrinsic nature of the classes, etc.

- From ML perspective, speech is a quite unique problem due to the nature of the input and class label outputs:
  - About the input → Time sequence
    - Very different length of the input wrt. output → Segmentation problem
    - Elasticity of the temporal dimension
    - Discriminative cues often distributed over a long temporal span
  - About the output → Output can be a sequence of class labels
    - Too much combinations → **Need structure!!!**

Speech Pattern Classification
The "simple" task

- The "simple" SPC task:

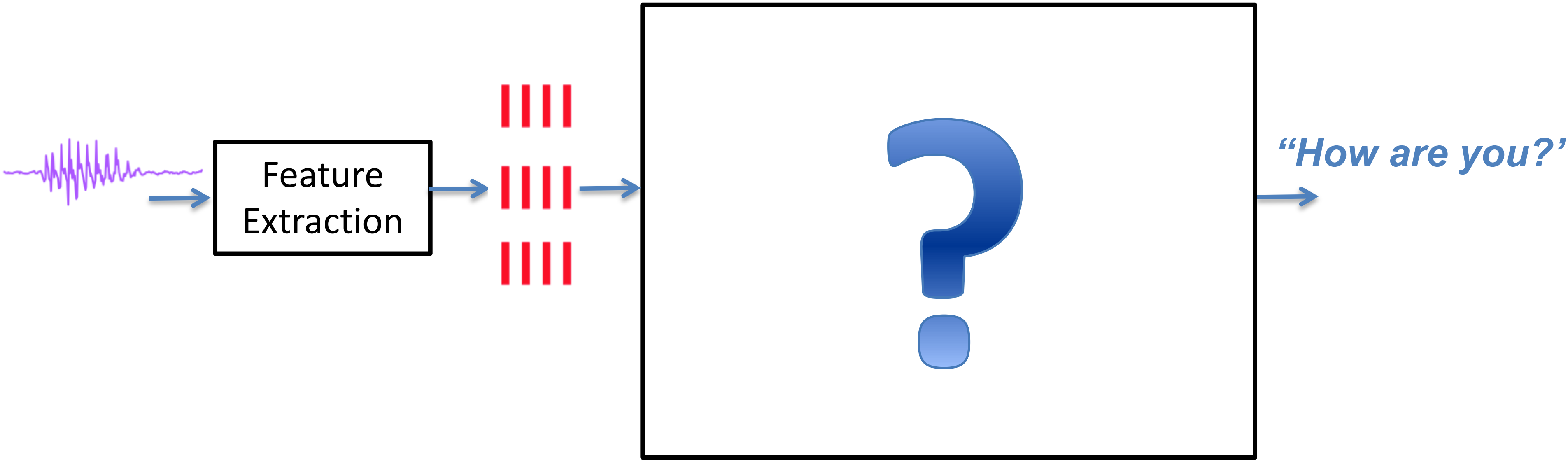


  - Static output:
    - No sequence of output labels
    - No segmentation problem → Audio segment corresponds to single class
  - No structured knowledge → Models correspond to output labels
  - Notice that:
    - Although being "simpler" from the ML perspective, they can be very hard
    - Can be classification/identification, verification or regression problems
    - Time-varying input still needs to be addressed

Speech Pattern Classification

The "complex" task

- **Goal** Given a sequence of observations determine which is the most likely sequence of words

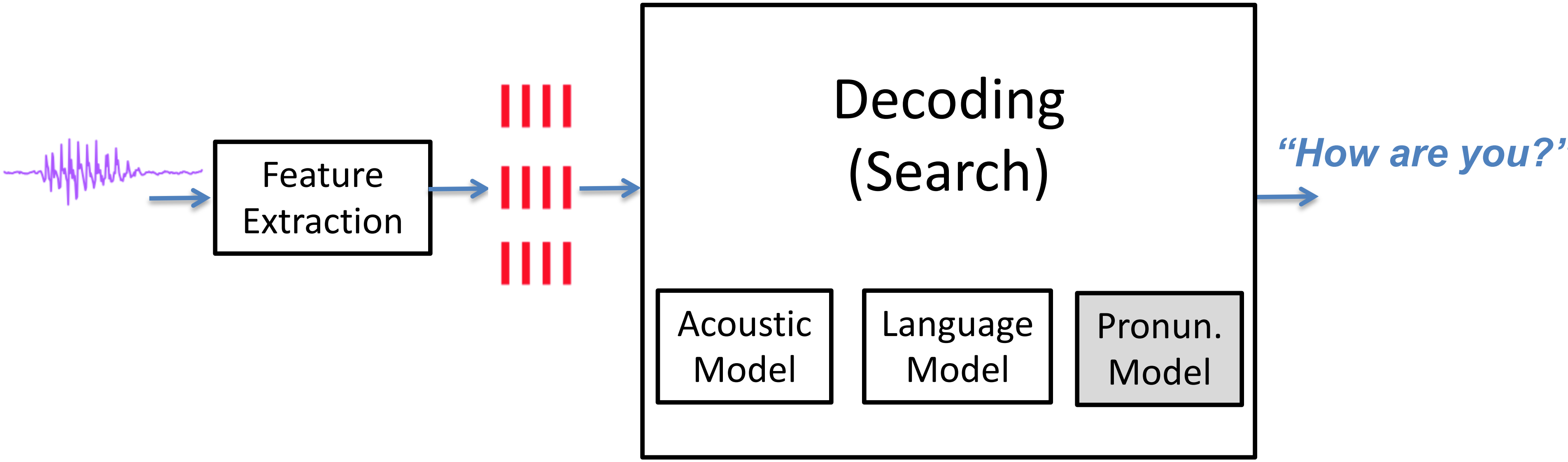


- Already decades of research on ASR (and other SLT related topics) → **Very challenging!!!**
- **Related sub-tasks:** Isolated ASR, Continuous ASR, KWS, LVCSR, STD/Search on Speech, etc.

Speech Pattern Classification

The “complex“ task

- **Goal** Given a sequence of observations determine which is the most likely sequence of words

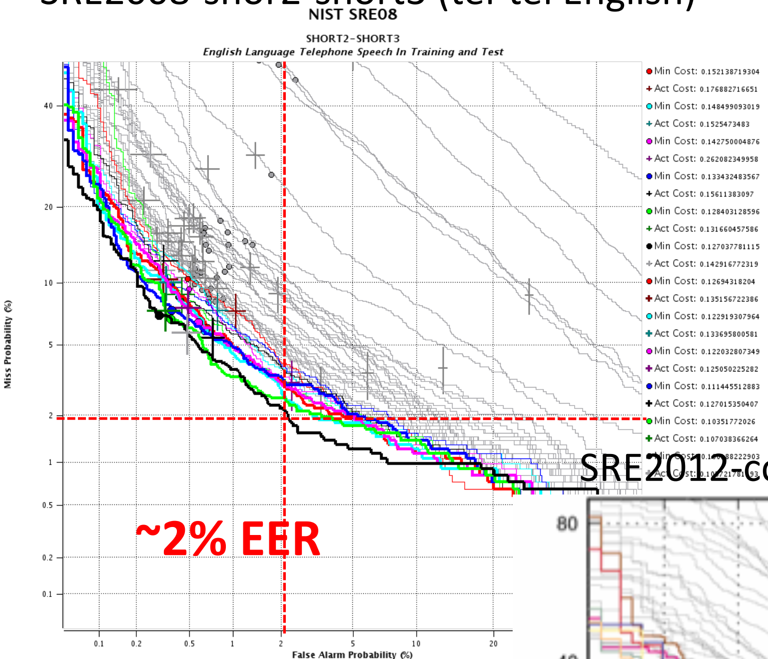


- Already decades of research on ASR (and other SLT related topics)
  → **Very challenging!!!**
- **Related sub-tasks:** Isolated ASR, Continuous ASR, KWS, LVCSR, STD/Search on Speech, etc.

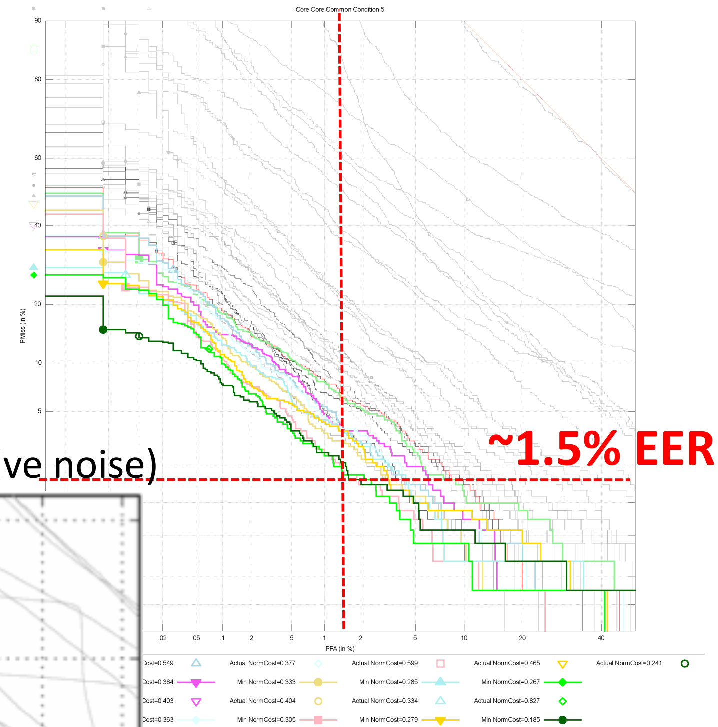
Speech Pattern Classification

How mature are these technologies? NIST SRE evolution

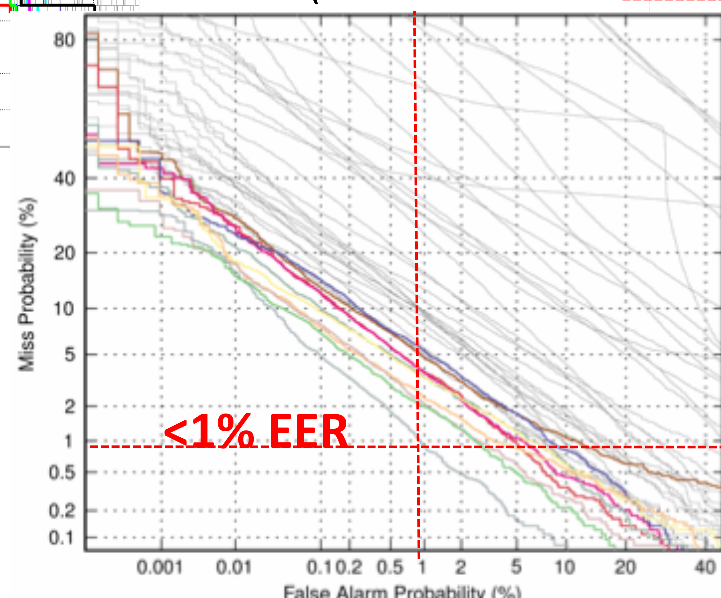
SRE2008-shor2-short3 (tel-tel English)



SRE2010-cc5 (tel-tel normal vocal effort)

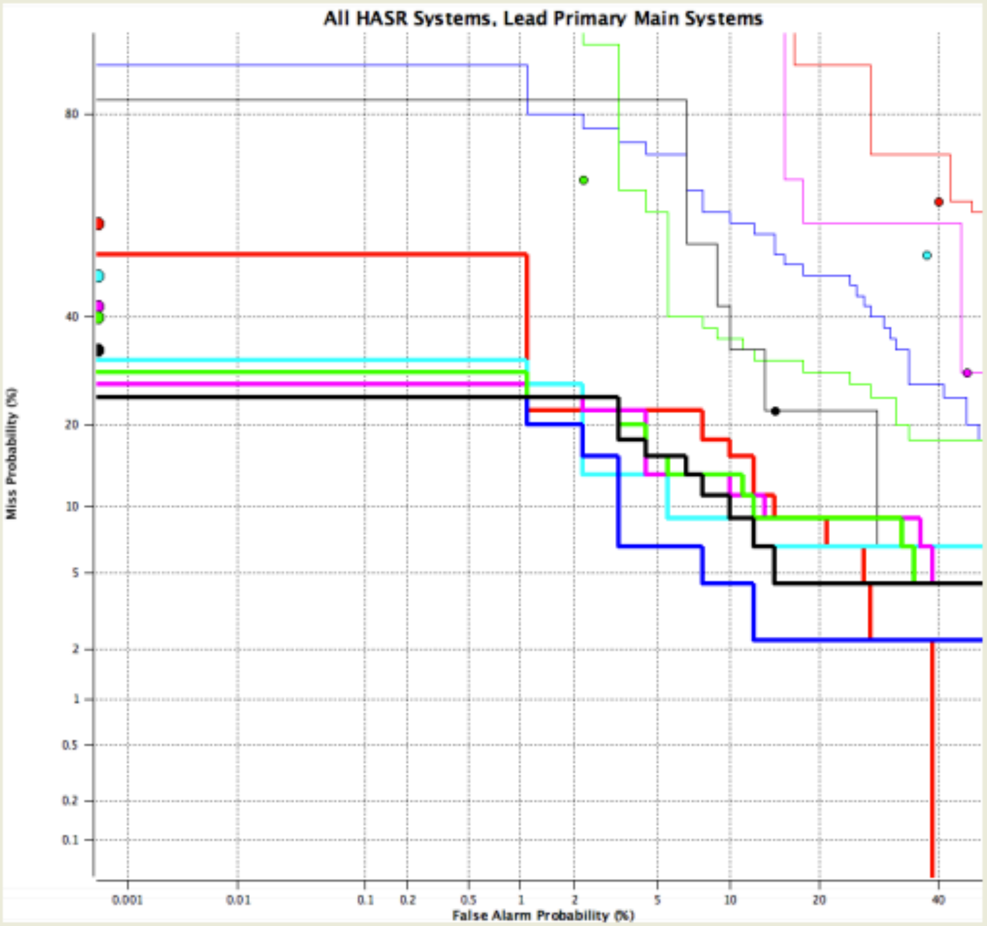


SRE2012-cc2 (tel-tel no additive noise)



Speech Pattern Classification

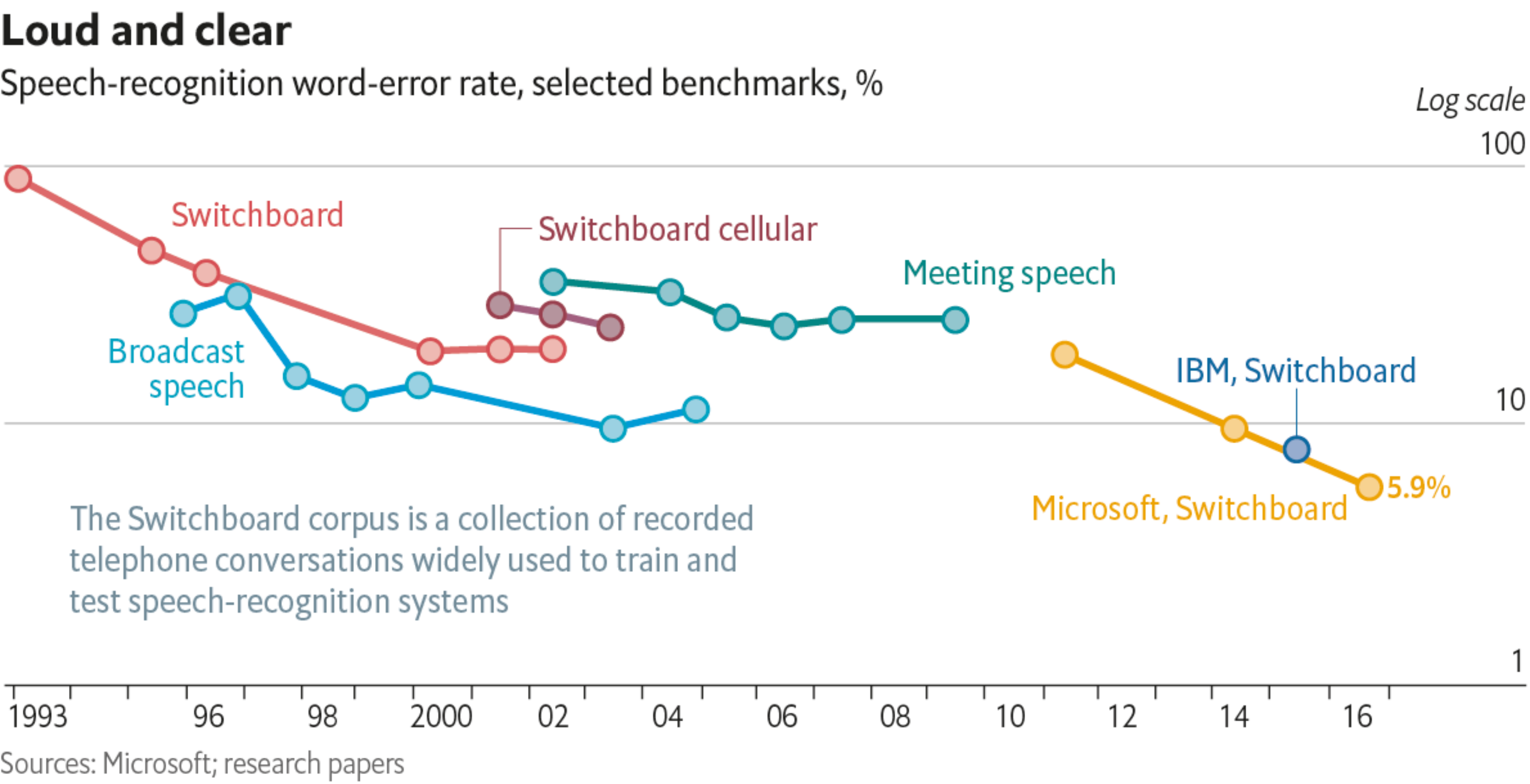
How mature are these technologies? NIST HASR 2010 results



- 135 HASR2 trials
- Six HASR systems (thin lines)
  one system = decision only
- Six Automatic systems (thick lines)

Speech Pattern Classification

How mature are these technologies? ASR Benchmark history



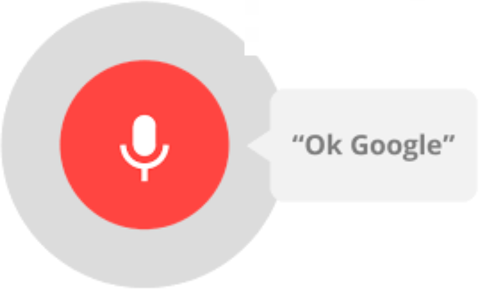
Speech Pattern Classification

How mature are these technologies? Industry

amazon alexa

DRAGON NATURALLYSPEAKING

NUANCE



Baidu Research

Outline

- Introduction to speech processing
- Speech pattern classification
- **Selected research topics**
- Two recent research works:
  - Domain adaptation for low-resource ASR
  - Native language (L1) identification

Selected Research Topics

Core application areas @ L^{2}F

Core SPEECH+NLP technologies

Multimedia transcription / analytics

- Rich transcription (1)
- Robust speech processing
- Speaker characterization (2)
- Speech analytics/Text mining
- Audio & Music processing

voice interaction

Speech-to-speech translation

- MT4 Microblogs
- Voice conversion
- Quality evaluation

Unbabel

Spoken/multimodal dialog systems

- Smart homes
- Cultural heritage
- Chat bots

E-Health

- Virtual therapists (3)
- Monitoring neurodeg.
- Detection of pathological speech, depression, etc.
- Active ageing
- IE/IR for medical texts
- Sign language

Other emerging areas

- Security/Privacy (4)
- Cyberbullying
- Brain interfaces

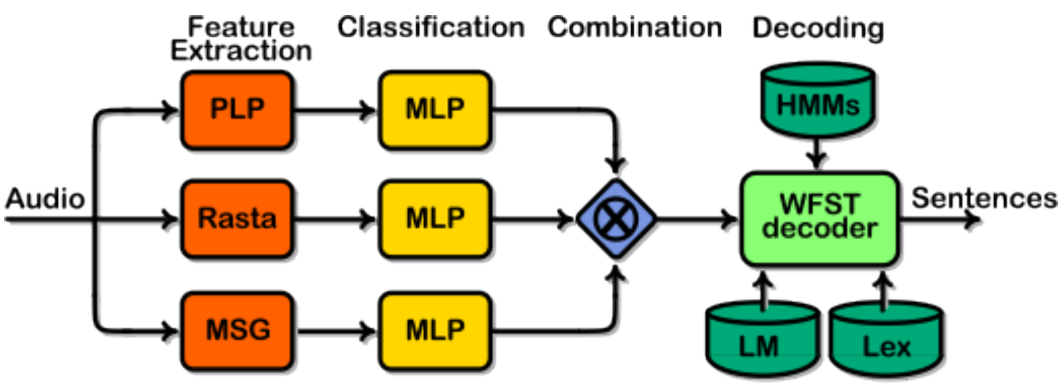
E-Learning

- Language learning
- Language acquisition

Selected Research Topics

TOPIC1 Multi-media and robust spoken language processing

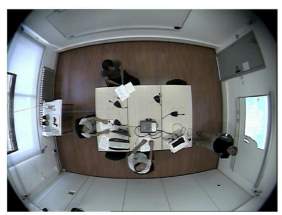
- **Large vocabulary ASR**
  - BN and multimedia transcription
  - Conversational telephone speech
  - Improved acoustic modeling
  - Language & dialect adaptation
  - Multi-dialectal ASR (PT and ES)
  - **Domain adaptation for low-resource languages**

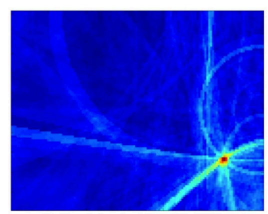


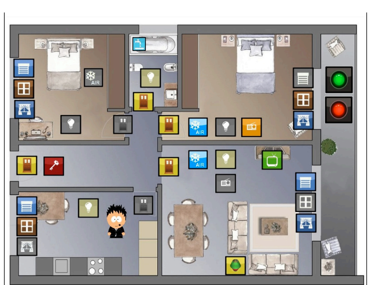
Selected Research Topics

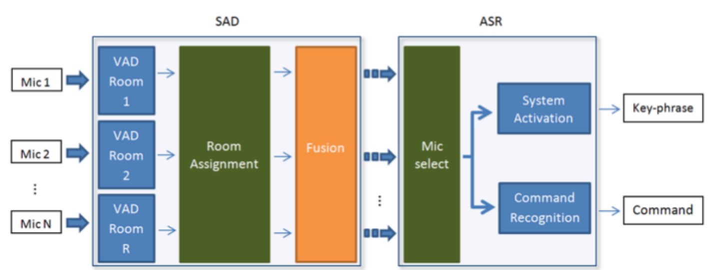
TOPIC1 Multi-media and robust spoken language processing

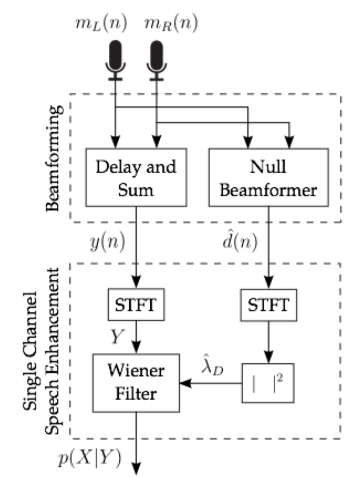
- **Robust distant speech processing for ASR**
  - Meeting assistance, home automation & robotic applications
  - Multi-room equipment and data collection of PT voice commands
  - Acoustic segmentation and diarization (in multi-room)
  - Multi-channel processing, beamforming and channel selection
  - Integration of beamforming and uncertainty propagation
  - Uncertainty propagation for DNN and dynamic model adaptation
  - <u>Challenges Participation</u>: CLEAR 2006, 2007, Pascal-CHIME 2011, Evalita 2014







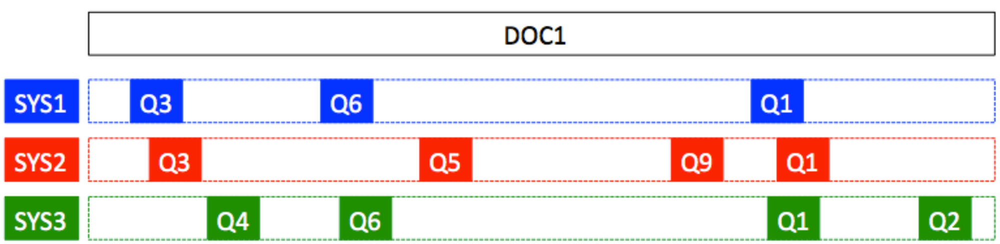


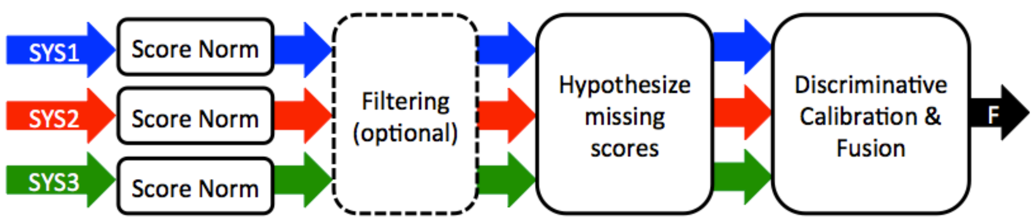


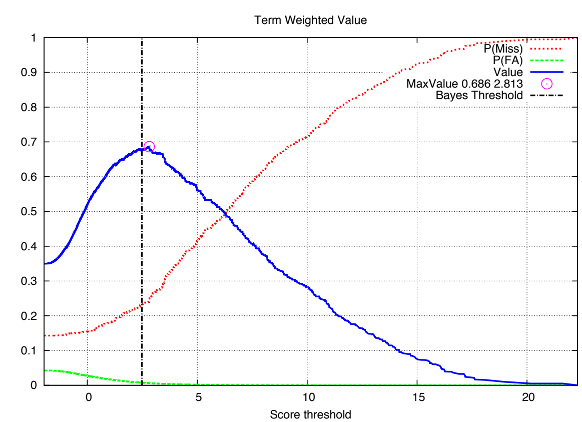
Selected Research Topics

TOPIC1 Multi-media and robust spoken language processing

- **Other topics**
  - Classification of semantic audio events in VIDIVIDEO
  - Key-word spotting for DNN/HMM ASR systems
  - Search on speech (big-data):
    - AKWS and DTW systems
    - Calibration and Fusion of heterogeneous (STD) detectors
    - <u>Challenges Participation</u>: Mediaeval 2012, 2013, Albayzin2016



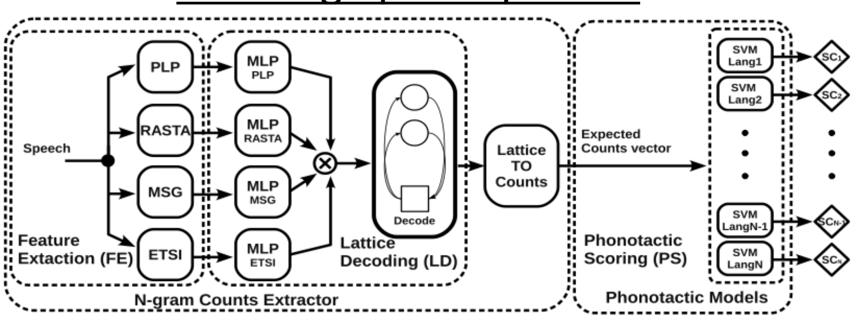


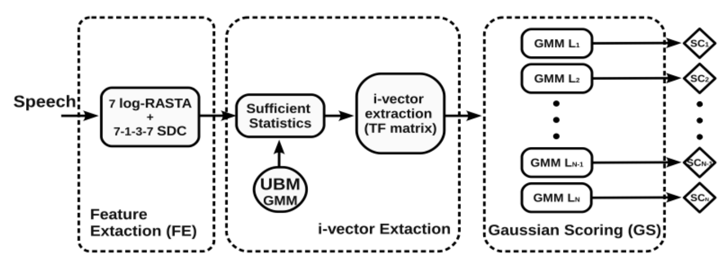


Selected Research Topics

TOPIC2 Speaker characterization

- **Language and dialect**
  - Phonotactic & i-vector systems
  - Integration in euTV multi-media processing pipeline
  - Challenge participation: Albayzin 2008, 2010, 2012 & NIST LRE 2009 & 2011;
- **Accent and nativeness**
  - Portuguese variety identification
  - English Nativeness detection (TED talks)
  - Degree of Nativeness regression based on multiple system fusion
  - **L1 Native Language identification of English students**
  - Challenge participation: ComParE 2015-2016, NLI shared task 2017;

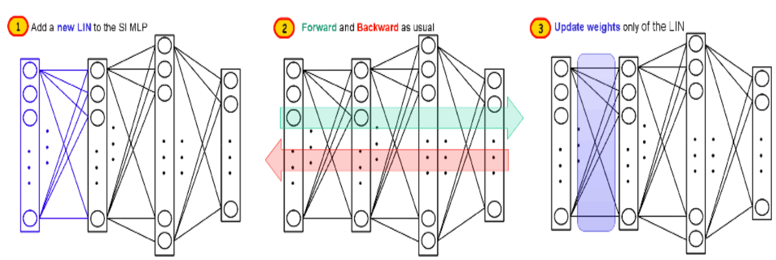


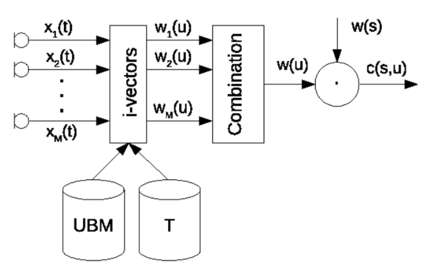


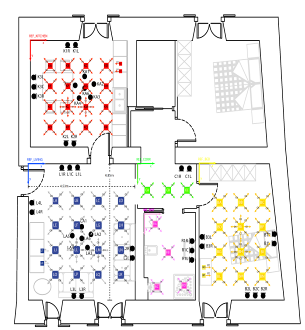
Selected Research Topics

TOPIC2 Speaker characterization

- **Speaker Identity**
  - Vector-based methods (i-vector)
  - Robust high-level features for speaker ID
  - Domestic speaker with multiple distant microphones
  - Challenge participation: NIST SRE 2010, SRE in Mobile Environments 2013
- **Other (paralinguistics)**
  - Gender & age → Detection of child voices in IDASH
  - Speech analytics for IVR monitorization: age & gender, dialogue hotspots detection,

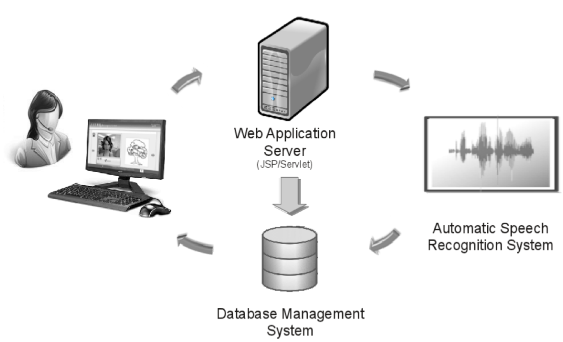


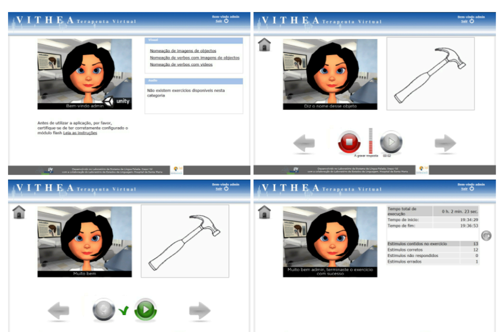


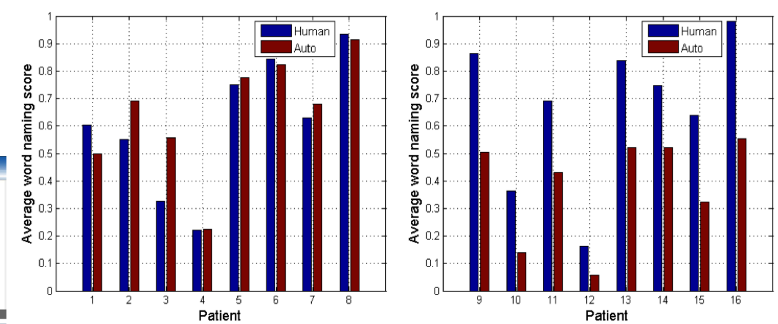


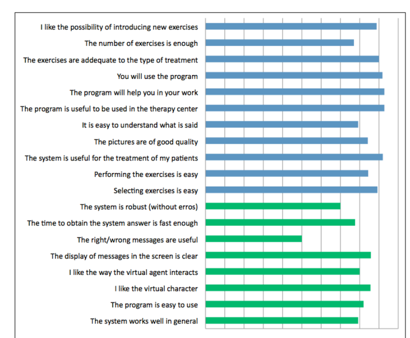
Selected Research Topics
TOPIC3 Health applications

- Speech and Language Technologies applied to **therapy**
  - VITHEA project → Virtual Therapy for **APHASIA**
    - Collection of aphasic speech corpus
    - Research on word verification for automatic exercise evaluations
    - Development and evaluation of virtual therapist prototype (web-based)
    - Several hospitals and therapists have used (are using) the system
    - Awarded with several prizes





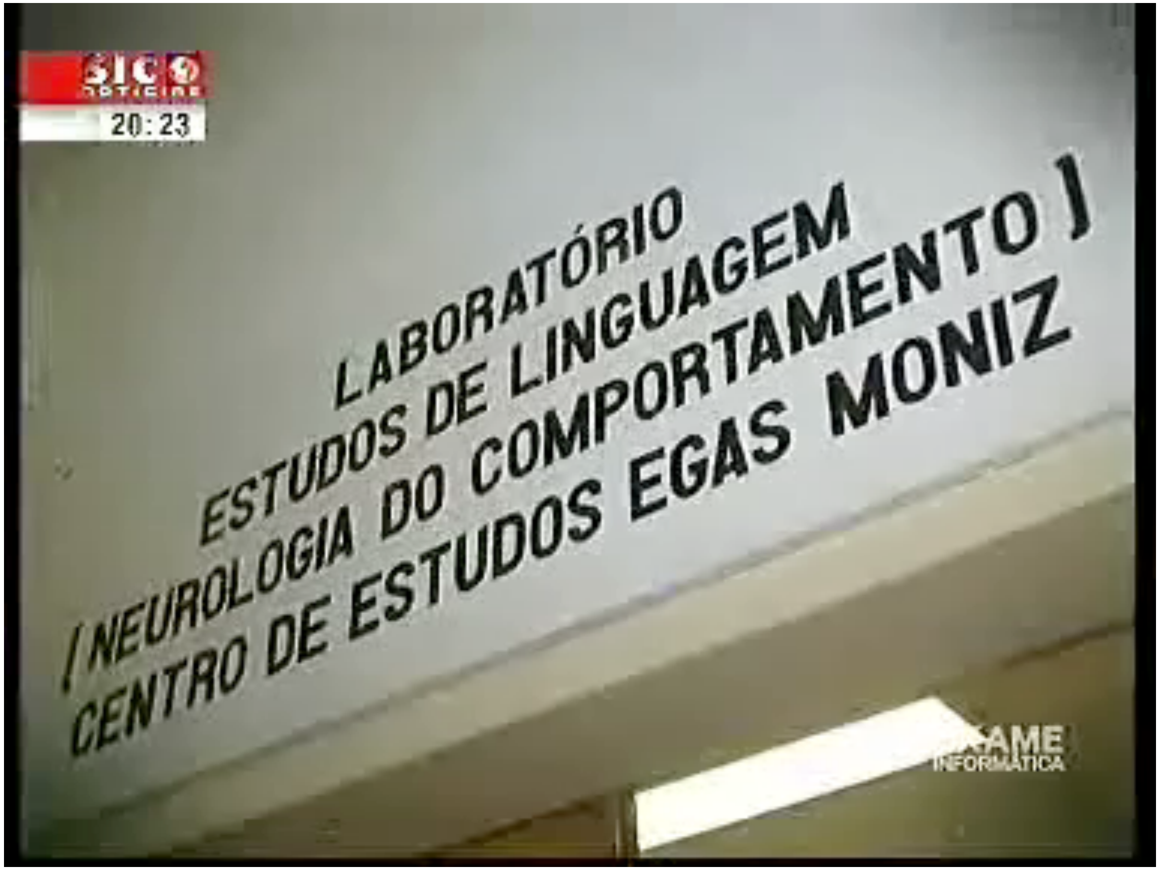




Selected Research Topics

TOPIC3 Health applications

- Speech and Language Technologies applied to **therapy**



Selected Research Topics

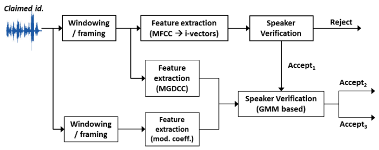
TOPIC3 Health applications

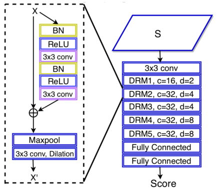
- Speech and Language Technologies applied to **diagnoses**
  - Cognitive Impairment Screening tool
    - Development and evaluation of on-line platform of neuropsychological tests; focus on verbal fluency tasks
  - Parkinson detection
    - Study of features for automatic detection on different tasks
    - Detecting dug on/off state
  - Dementia/Alzheimer detection
    - Exploiting pragmatic features → Topic coherence
  - Children Pathological speech
    - Children articulation and language disorders
    - Data collection; Improvement of ASR models; gamification
  - Apnea and sleep disorders
    - Speech as a Biomarker for Obstructive Sleep Apnea

Selected Research Topics

TOPIC4 Privacy and security

- Privacy-preserving speech processing
  - Investigation on algorithms, protocols, methods for speech processing "without access to speech"
    - Music-matching algorithms with secure multi-party (SMC) protocols
    - Speaker verification → SBE + i-vectors, GMM Garbled Circuits
      (no access to speech access, neither to speaker model)
    - Document retrieval (speech search) → SBE + DTW
    - Paralinguistic classification with DNNs (Cryptonets)
  - Application of de-identification to sensible tasks:
    - Depression detection on de-identified speech
- Security in speaker authentication:
  - Spoofing attack detection:
    - Automatic detection of converted speech
    - Automatic detection of replay attacks





Outline

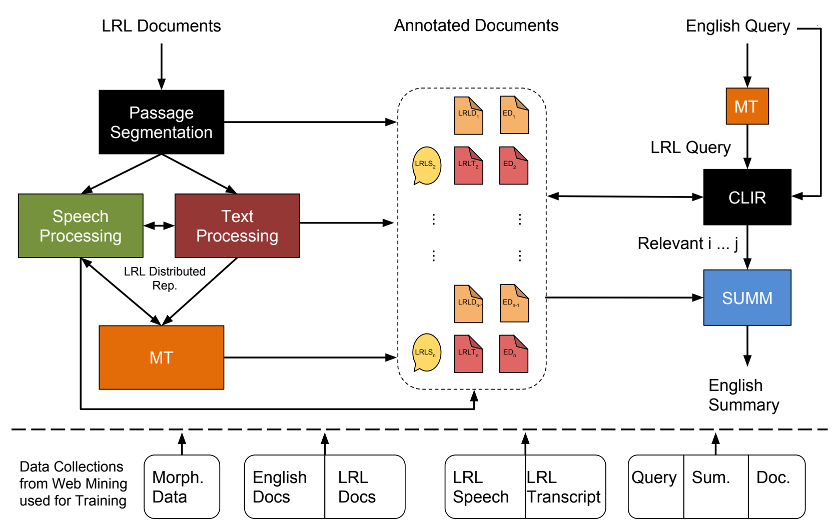
- Introduction to speech processing
- Speech pattern classification
- Selected research topics
- **Two recent research work examples:**
  - **Domain adaptation for low-resource ASR**
  - Native language (L1) identification

Domain adaptation for low-resource ASR

Last year during my sabbatical @ Univ. Edinburgh

- MATERIAL programme seeks to develop methods for searching speech and text in **low-resource** languages using English queries
- ASR systems must operate on diverse multi-genre data, including **telephone conversations**, **news** and **topical broadcasts**
- The only manually annotated training data is from the telephone speech domain

MATERIAL

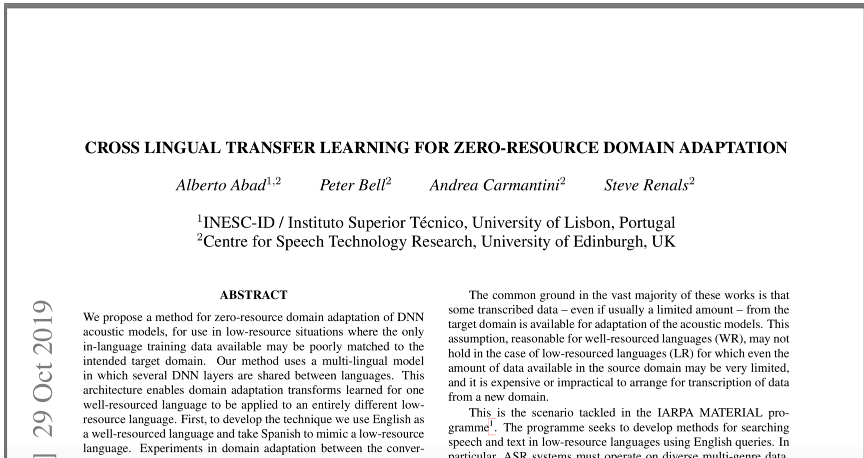


Domain adaptation for low-resource ASR

Objectives

- **Goal:** Transfer specific channel/style conditions learnt in a well-resourced (WR) language to a low-resourced (LR) language for which training data is not available
- **How?**
  - Train multi-lingual/multi-task AM with WR+LR data in a common channel/style (ie. CTS).
  - Adapt network using new channel/style WR data (ie. BN):
    - Adapt only (at most) up to the last common layer, so the last language specific layers an unchanged.
  - Transfer adapted first layer weights and concatenate with LR last layers.
- Related with transfer learning, model adaptation, low resource ASR, multi-lingual learning, etc.

Abad et al. (2019), "Cross lingual transfer learning for zero-resource domain adaptation", https://arxiv.org/abs/1910.02168



CROSS LINGUAL TRANSFER LEARNING FOR ZERO-RESOURCE DOMAIN ADAPTATION

Alberto Abad[1,2] *Peter Bell*[2] *Andrea Carmantini*[2] *Steve Renals*[2]

[1]INESC-ID / Instituto Superior Técnico, University of Lisbon, Portugal
[2]Centre for Speech Technology Research, University of Edinburgh, UK

29 Oct 2019

ABSTRACT

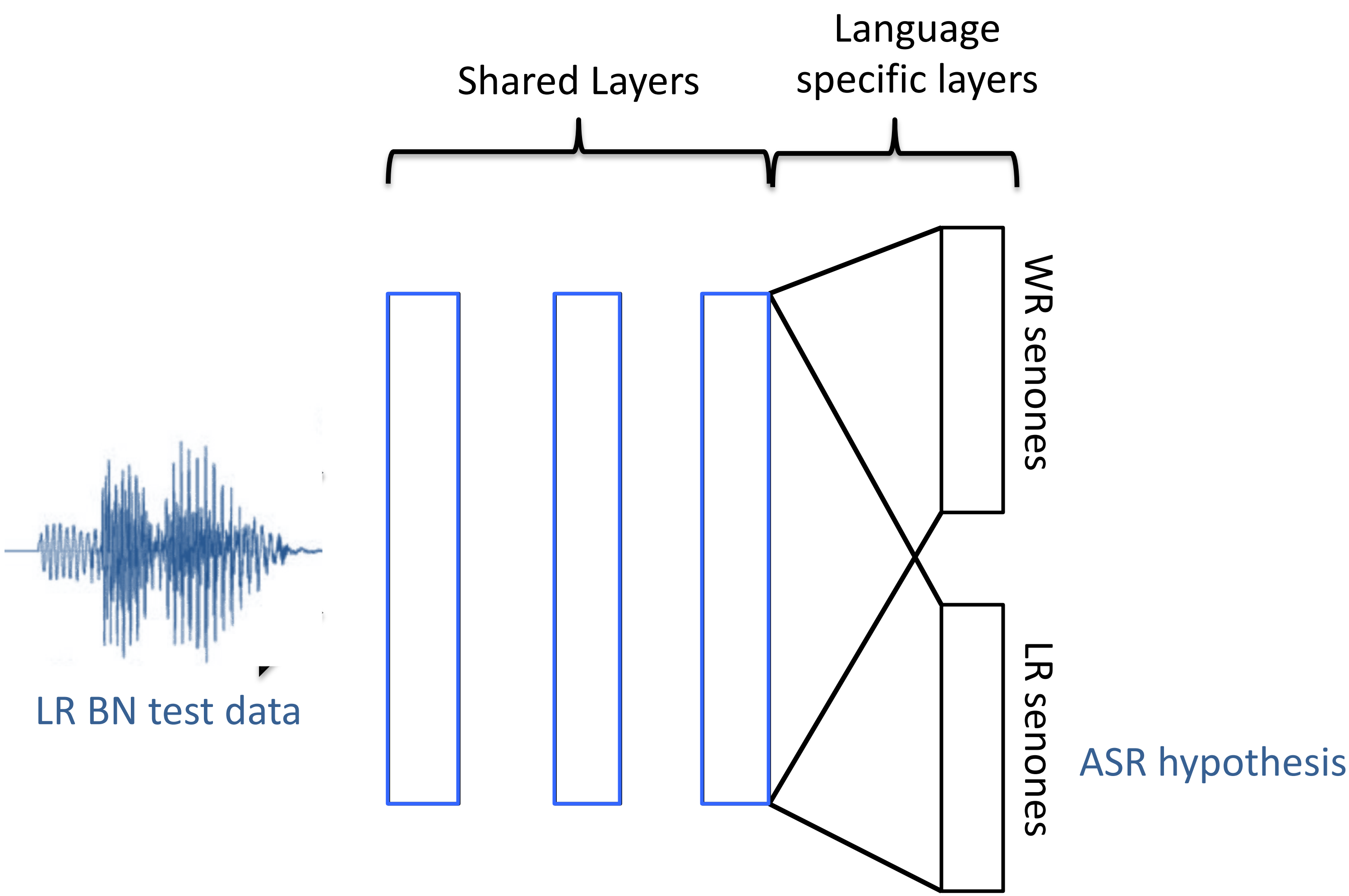
We propose a method for zero-resource domain adaptation of DNN acoustic models, for use in low-resource situations where the only in-language training data available may be poorly matched to the intended target domain. Our method uses a multi-lingual model in which several DNN layers are shared between languages. This architecture enables domain adaptation transforms learned for one well-resourced language to be applied to an entirely different low-resource language. First, to develop the technique we use English as a well-resourced language and take Spanish to mimic a low-resource language. Experiments in domain adaptation between the conver-

The common ground in the vast majority of these works is that some transcribed data – even if usually a limited amount – from the target domain is available for adaptation of the acoustic models. This assumption, reasonable for well-resourced languages (WR), may not hold in the case of low-resourced languages (LR) for which even the amount of data available in the source domain may be very limited, and it is expensive or impractical to arrange for transcription of data from a new domain.

This is the scenario tackled in the IARPA MATERIAL programme[1]. The programme seeks to develop methods for searching speech and text in low-resource languages using English queries. In

Domain adaptation for low-resource ASR

Cross-lingual adaptation approach



Domain adaptation for low-resource ASR

Baseline and multi-task results

	Test condition			
	WR		LR	
	CTS source	BN target	CTS source	BN target
mono-ling BN AM	---	11.8	---	19.2
mono-ling CTS AM	22.6	19.6	32.3	40.0
multi-ling CTS AM	23.6	19.2	32.6	32.9

- CTS (Fisher) is the source condition and BN (hub4) is the target condition
- Spanish is the LR language and English the WR language
- Experimental set-up:
  - TDNN hires + pitch, no LF-MMI, no ivecs, all downsampled to 8kHz
  - Use of matched LMs (CTS/BN test data is decoded with CTS/BN LM)

Domain adaptation for low-resource ASR

Proposed cross-lingual adaptation results

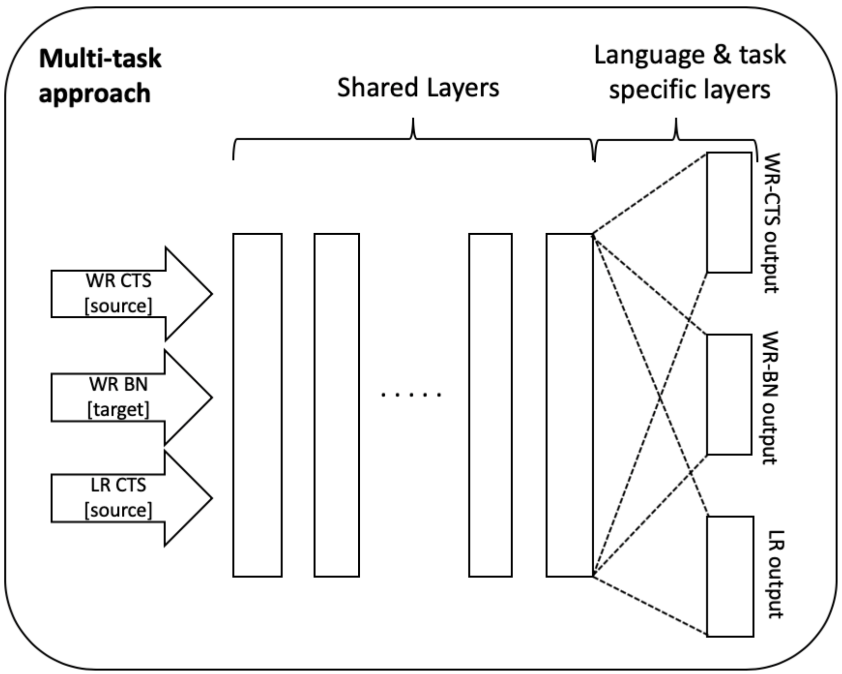
	LR language	WR language
	BN	BN
mono-ling CTS AM	40.0	19.6
multi-ling CTS AM	32.9	19.2
proposed CL adapt AM	**28.4**	14.5
Upper-bound (BN training)	19.2	11.8

- From 40.0% to 32.9% thanks to multilang and from 32.9% to 28.4 to nnet adapt & transfer learning → NO USE OF ANY ADDITIONAL LR TRAINING DATA!!!

Domain adaptation for low-resource ASR

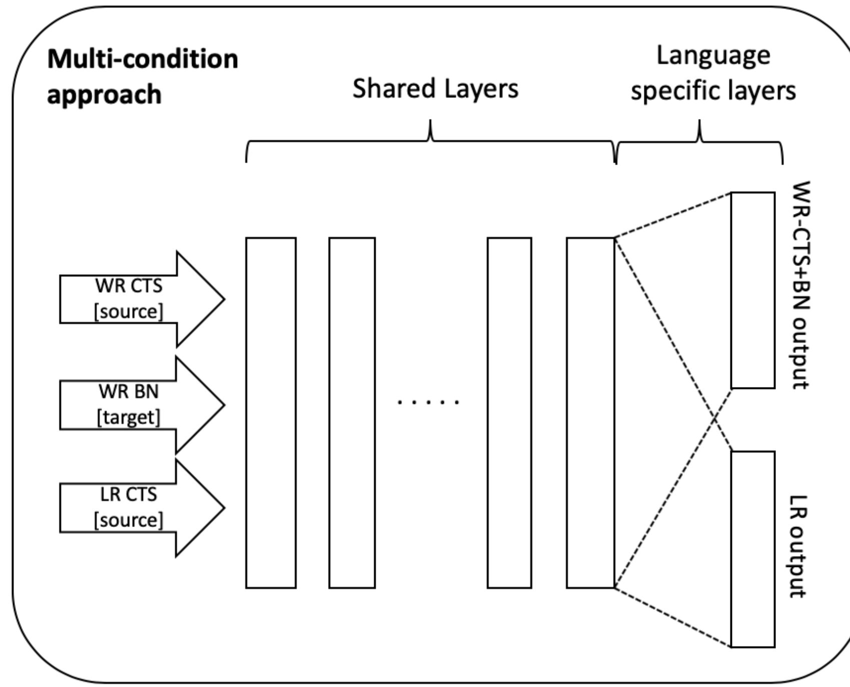
Comparison with similar approaches

Multi-task approach



- Train in a multi-task way a nnet with 3 language-task pairs
- Use the LR output for decoding target BN data

Multi-condition approach



- Train in a multi-condition way a nnet with 2 lang outputs (WR data is mixed)
- Use the LR output for decoding target BN data

Domain adaptation for low-resource ASR

Comparison with similar approaches

	LR language	WR language
	BN	BN
mono-ling CTS AM	40.0	19.6
multi-ling CTS AM	32.9	19.2
proposed CL adapt AM	**28.4**	14.5
multitask	29.1	12.4
multitask + fine-tuning	**29.1**	12.3
multicondition	29.2	12.5
multicondition + fine-tuning	**29.1**	12.2
Upper-bound (BN training)	19.2	11.8

- Experiments adapting 1 to 3 first layers for 0.5 and 1 epochs → Results show the best configuration for each case.
- For multilang/multicondition, there is not a noticeable improvement for LR in-domain ASR:
  - Performances for the different adaptation parameters oscillate in +/-0.1 differences.
  - Best results are obtained with the minimum amount of adaptation (not able to further exploit WR-BN data)

Domain adaptation for low-resource ASR

Experiments with low-resourced languages

	Tagalog			Lithuanian		
	wideband data			wideband data		
	NB	TB	avg	NB	TB	avg
baseline CTS	53.2	58.7	57.3	45.6	43.0	44.0
multilang CTS	46.5	52.2	50.7	38.2	36.5	37.1
proposed CL adapt AM	41.9	48.5	**46.8**	31.6	32.1	**31.9**

- Use BABEL training set:
  - Exact same architecture as previous experiments
- Eval on BABEL dev and Material *analysis_** test sets:
  - CSTR MATERIAL LM → Trained on *webnews*

Outline

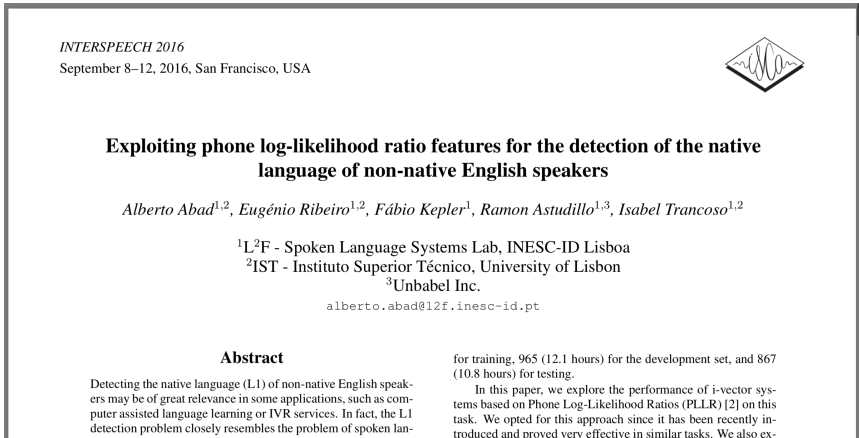
- Introduction to speech processing
- Speech pattern classification
- Selected research topics
- **Two recent research work examples:**
  - Domain adaptation for low-resource ASR
  - **Native language (L1) identification**

Native Language (L1) identification

Objectives

- The **ComParE 2016 Native Language** task aims at identifying L1 of non-native English speakers:
  - Similar to language, accent, and dialect ID in Spoken Language Recognition (SLR)
    - Most successful systems are based on acoustic or phonotactic information
    - Combination tends to provide increased performance
    - **Phone Log-Likelihood Ratio (PLLR)** features convey frame-by-frame acoustic-phonetic information.
- The main **objective** is to explore PLLR features in the L1 detection task, and also:
  - Comparison of PLLR with acoustic and phonotactic approaches
  - Use of (as much as possible) in-house already available technology
  - Develop a (hopefully) good performing system and have fun!!

Abad et al. (2016), "Exploiting Phone Log-Likelihood Ratio Features for the Detection of the Native Language of Non-Native English Speakers", in Proc. Interspeech 2016
https://www.inesc-id.pt/publications/12183/pdf


INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Exploiting phone log-likelihood ratio features for the detection of the native language of non-native English speakers

Alberto Abad[1,2], Eugénio Ribeiro[1,2], Fábio Kepler[1], Ramon Astudillo[1,3], Isabel Trancoso[1,2]

[1]L^{2}F - Spoken Language Systems Lab, INESC-ID Lisboa
[2]IST - Instituto Superior Técnico, University of Lisbon
[3]Unbabel Inc.
alberto.abad@l2f.inesc-id.pt

Abstract

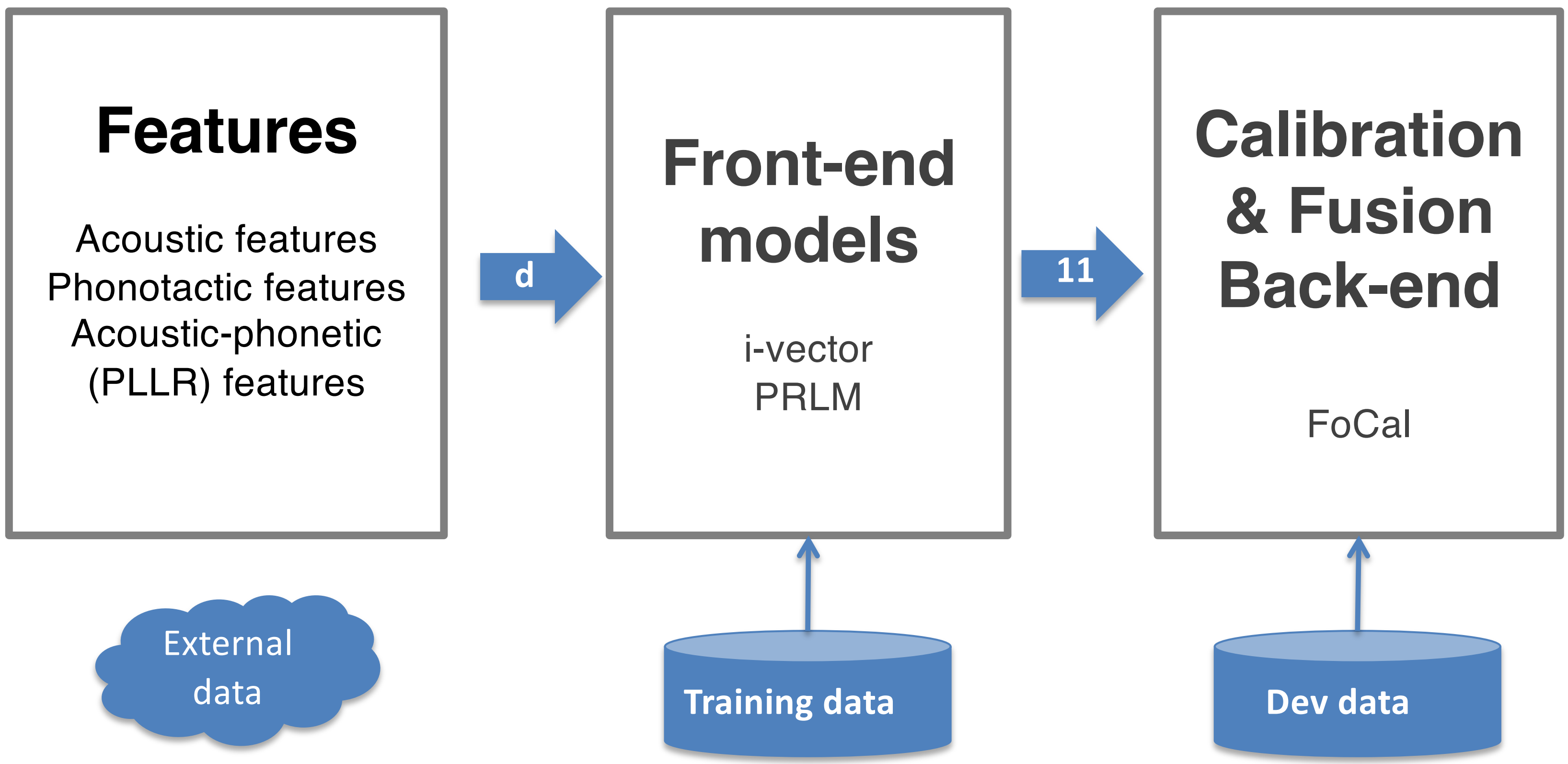
Detecting the native language (L1) of non-native English speakers may be of great relevance in some applications, such as computer assisted language learning or IVR services. In fact, the L1 detection problem closely resembles the problem of spoken lan-

for training, 965 (12.1 hours) for the development set, and 867 (10.8 hours) for testing.

In this paper, we explore the performance of i-vector systems based on Phone Log-Likelihood Ratios (PLLR) [2] on this task. We opted for this approach since it has been recently introduced and proved very effective in similar tasks. We also ex-

Native Language (L1) identification

INESC-ID approaches for L1 identification



Native Language (L1) identification

Results in the DEV set

	UAR [%]	Acc [%]
Baseline	45.1	44.9
Phonotactic (BR)	46.4	46.2
Phonotactic (EN)	51.4	51.4
Phonotactic (ES)	50.0	49.8
Phonotactic (PT)	53.1	53.1
Phonotactic (ALL) (I)	63.3	63.2
i-vectors (MFCC) (II)	76.2	76.3
i-vectors (BR-PLLR)	76.9	76.9
i-vectors (EN-PLLR)	79.2	79.2
i-vectors (ES-PLLR)	77.6	77.4
i-vectors (PT-PLLR)	80.6	80.5
i-vectors (ALL PLLR) (III)	83.0	82.9
(I) + (II)	78.6	78.7
(II) + (III)	84.6	84.6

Native Language (L1) identification

Results in the DEV set

ComPaRe 2016 Official Baseline

	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR
ARA	29	3	5	7	5	5	6	6	7	6	7
CHI	4	38	5	4	5	2	5	10	6	4	1
FRE	11	7	29	8	0	4	3	1	11	0	6
GER	5	3	5	55	1	7	1	2	5	1	0
HIN	4	1	1	0	47	2	2	2	2	21	1
ITA	6	2	9	6	6	46	0	4	10	1	4
JPN	4	13	4	2	2	1	36	11	10	1	1
KOR	4	19	1	2	2	3	14	32	5	3	5
SPA	6	11	15	6	2	4	9	9	32	1	5
TEL	2	0	2	2	24	2	2	2	2	43	2
TUR	6	5	5	5	2	6	7	8	5	0	46

INESC-ID ComPaRe 2016 system

	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR
ARA	77	0	3	1	0	1	1	0	1	0	2
CHI	0	78	0	1	0	1	2	0	1	1	0
FRE	3	0	64	2	0	2	2	0	5	0	2
GER	2	1	2	78	0	0	0	1	0	0	1
HIN	0	0	0	0	67	0	0	0	0	16	0
ITA	1	0	5	2	0	79	1	1	3	0	2
JPN	1	1	1	0	0	0	70	8	4	0	0
KOR	2	4	1	1	0	0	5	77	1	0	0
SPA	2	1	2	1	0	5	4	5	77	1	2
TEL	0	0	0	0	18	0	0	0	0	65	0
TUR	0	1	1	3	1	2	0	2	1	0	84

Native Language (L1) identification

Final results in the TEST set

	DEV [UAR %]	TEST [UAR %]
ComPaRe 2016 Official Baseline	45.1%	47.5%
INESC-ID ComPaRe 2016 system	84.6%	81.3%

Native Language (L1) identification

Final results in the TEST set

Not Secure — compare.openaudio.eu

Winners of the INTERSPEECH 2016 Computational Paralinguistics Challenge:

- ***The Deception Sub-Challenge is awarded to:***
  CLAUDE MONTACIÉ, MARIE-JOSÉ CARATY:
  Prosodic Cues and Answer Type Detection for the Deception Sub-Challenge
- ***The Sincerity Sub-Challenge is awarded to:***
  HEYSEM KAYA, ALEXEY A. KARPOV:
  Fusing Acoustic Feature Representations for Computational Paralinguistics Tasks
- ***The Native Language Sub-Challenge is awarded to:***
  ALBERTO ABAD, EUGÉNIO RIBEIRO, FÁBIO KEPLER, RAMON ASTUDILLO, ISABEL TRANCOSO:
  Exploiting Phone Log-likelihood Ratio Features for the Detection of the Native Language of Non-native English Speakers

Winners of the INTERSPEECH 2015 Computational Paralinguistics Challenge:

- ***The Degree of Nativeness Sub-Challenge Prize is awarded to:***
  MATTHEW P. BLACK, DANIEL BONE, ZISIS I. SKORDILIS, RAHUL GUPTA, WEI XIA, PAVLOS PAPADOPOULOS, SANDEEP NALLAN CHAKRAVARTHULA, BO XIAO, MAARTEN VAN SEGBROECK, JANGWON KIM, PANAYIOTIS G. GEORGIOU, SHRIKANTH S. NARAYANAN
  Automated Evaluation of Non-native English Pronunciation Quality: Combining Knowledge- and Data-driven Features at Multiple Time
- ***The Parkinson's Condition Sub-Challenge Prize is awarded to:***
  TAMÁS GRÓSZ, RÓBERT BUSA-FEKETE, GÁBOR GOSZTOLYA, LÁSZLÓ TÓTH

Native Language (L1) identification

Quiz

1 2 3 4 5 6 7

Arabic French German Italian Japanese Mandarin Chinese Spanish

Summary

- Speech processing has been the focus of extensive research during the last decades.
- As a result, there is a significant amount of very successful technologies in the market, such as Automatic Speech Recognition (ASR).
- ASR is a particularly difficult case of *speech pattern classification* due to the sequence to sequence nature of the task and the variability of speech:
  - Nevertheless, impressive results are attained nowadays in part thanks to the very positive impact of deep learning.
- In general, speech processing is becoming mature enough to foresee novel areas of application.
- Still, there are many open research challenges and problems.

Thank you!!

Questions?

Alberto Abad – alberto.abad@tecnico.ulisboa.pt

References

- These are some recommended tutorial-like reading in the topic of ASR:

  - 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.
  - 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.
  - D. Yu and J. Li (2017) Recent progresses in deep learning based acoustic models, in *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 3, pp. 396-409, 2017.
  - R Prabhavalkar et al. (2017) A Comparison of Sequence-to-Sequence Models for Speech Recognition, in Proc. Interspeech 2017.

Tools for Feature Extraction: HTK

HTK **http://htk.eng.cam.ac.uk**

- HMM toolkit primarily used for ASR
  - I has been one of the most important publicly available ASR toolkits for many years
  - Provides source code written in C (Linux/Windows)
    - It does not allow re-distribution
  - Well-documented
- Contains several tools, including **HCopy** , the tool that allows for feature extraction
  - **HCopy** permits computation of the most relevant classical ASR features and typical pre-/post- processing:
    - LPC, FBE, MFCC, PLP
    - Energy, Delta, double-delta, CMVN, VTLN
  - It can read several audio input formats

Tools for Feature Extraction: openSMILE

openSMILE - Open-Source Audio Feature Extractor

SMILE - Speech & Music Interpretation by Large-space Extraction

http://audeering.com/research/opensmile/

- It is a extremely popular and versatile feature extraction tool in the area of paralinguistics:
  - Baseline in ComParE evaluations
- Open-source multi-platform (written in C++)
  - It permits stand-alone tool usage or library access
- Well-documented http://www.audeering.com/research-and-open-source/files/openSMILE-book-latest.pdf
- Popular I/O file formats are supported:
  - HTK, Comma separated value (CSV) text, WEKA, LibSVM

Tools for (speech) data modeling

GMM

- SPEAR: A Speaker Recognition Toolkit based on Bob (Python) https://pythonhosted.org/bob.bio.spear/
- MATLAB - Statistics and Machine Learning Toolbox http://www.mathworks.com/help/stats/fitgmdist.html

SVM

- LIBSVM -- A Library for Support Vector Machines https://www.csie.ntu.edu.tw/~cjlin/libsvm/

NEURAL NETWORKS

- Neural Network Toolbox http://www.mathworks.com/help/nnet/index.html
- QuickNet http://www1.icsi.berkeley.edu/Speech/qn.html

DNNs

- Theano, TensorFlow, CNTK, Keras, PyTorch

DATA MINING TOOLBOXES

- Weka 3: Data Mining Software in Java http://www.cs.waikato.ac.nz/ml/weka/
- SciKit learn (Python) http://scikit-learn.org/stable/

Tools for ASR development

HTK http://htk.eng.cam.ac.uk

- I has been one of the most important publicly available ASR toolkits for many years
- Provides source code written in C (Linux/Windows)
- Well-documented

KALDI http://kaldi-asr.org

- Provides current state of the art methods (DNNs)
- Many recipes ready to be used

Tools for LM training

- SRILM Toolkit: www.speech.sri.com/projects/srilm
- CMU-Cambridge Statistical LM toolkit: http://mi.eng.cam.ac.uk/~prc14/toolkit.html