EVALITA-ISTC Comparison of Open Source Tools on Clean and Noisy Digits Recognition Tasks

Piero Cosi
Mauro Nicolao

ISTITUTO DI SCIENZE E TECNOLOGIE DELLA COGNIZIONE
SEDE DI PADOVA - "FONETICA E DIALETTOLOGIA"
Consiglio Nazionale delle Ricerche
Via Martiri della libertà, 2 – 35137 Padova (Italy)
e-mail: {piero.cosi, mauro.nicolao}@pd.istc.cnr.it - www: http://www.pd.istc.cnr.it

EVALITA 2009

EVALITA 2009 - Connected Digits Recognition task
Reggio Emilia (Italia), 12 Dicembre 2009

Outline

- **Introduction**
  - **EVALITA speech tasks: connected digits recognition**
- **Data**
  - **Training, Development, Test**
- **ASR Architectures**
  - **CSLU Speech Toolkit**
  - **SONIC**
    - **Feature extraction**
      - **PMVDR: Perceptual Minimum Variance Distortionless Response**
  - **SPHINX**
- **Results**
- **Concluding Remarks**

Introduction

- Commissione di Gestione della base dei dati vocali italiani
  - Ministero delle Poste e Telecomunicazioni
  - 1991-1995
- ForumTAL
  - Coordinamento delle iniziative di ricerca e di sviluppo nel campo del Trattamento Automatico del Linguaggio
  - Ministero delle Comunicazioni
  - 2002
- EVALITA
  - Evaluation campaign of Natural Language Processing tools
  - AI*IA - NLP working group
  - 2007
- **EVALITA Speech Tasks**
  - **Connected Digits Recognition**
  - **Dialogue System Evaluation**
  - **Speaker Identity Verification (Application & Forensic)**
  - **AISV**
  - **2009**

nessuna campagna di valutazone sullo speech

Data

Sub Set	Clean Audio Files	Noisy Audio Files	Clean Digit Sequences	Noisy Digit Sequences
Training	3144	2204	10129	7376
Development	216	299	1629	1941
Test	365	605	2361	4036

0 [dz E r o]
1 [u n o]
2 [d u e]
3 [t r E]
4 [k w a t r o]
5 [tS i n k w e]
6 [s E I]
7 [s E t e]
8 [O t o]
9 [n O v e]

simple grammar
[<any> (<digit> [silence]) + <any>]

Within the EVALITA framework only the **orthographic transcriptions** are available so one of our previously-created general-purpose recognizer [9] has been used to create the **phonetically aligned transcriptions** needed from CSLU and SONIC systems to start the training.

CSLU Speech Toolkit

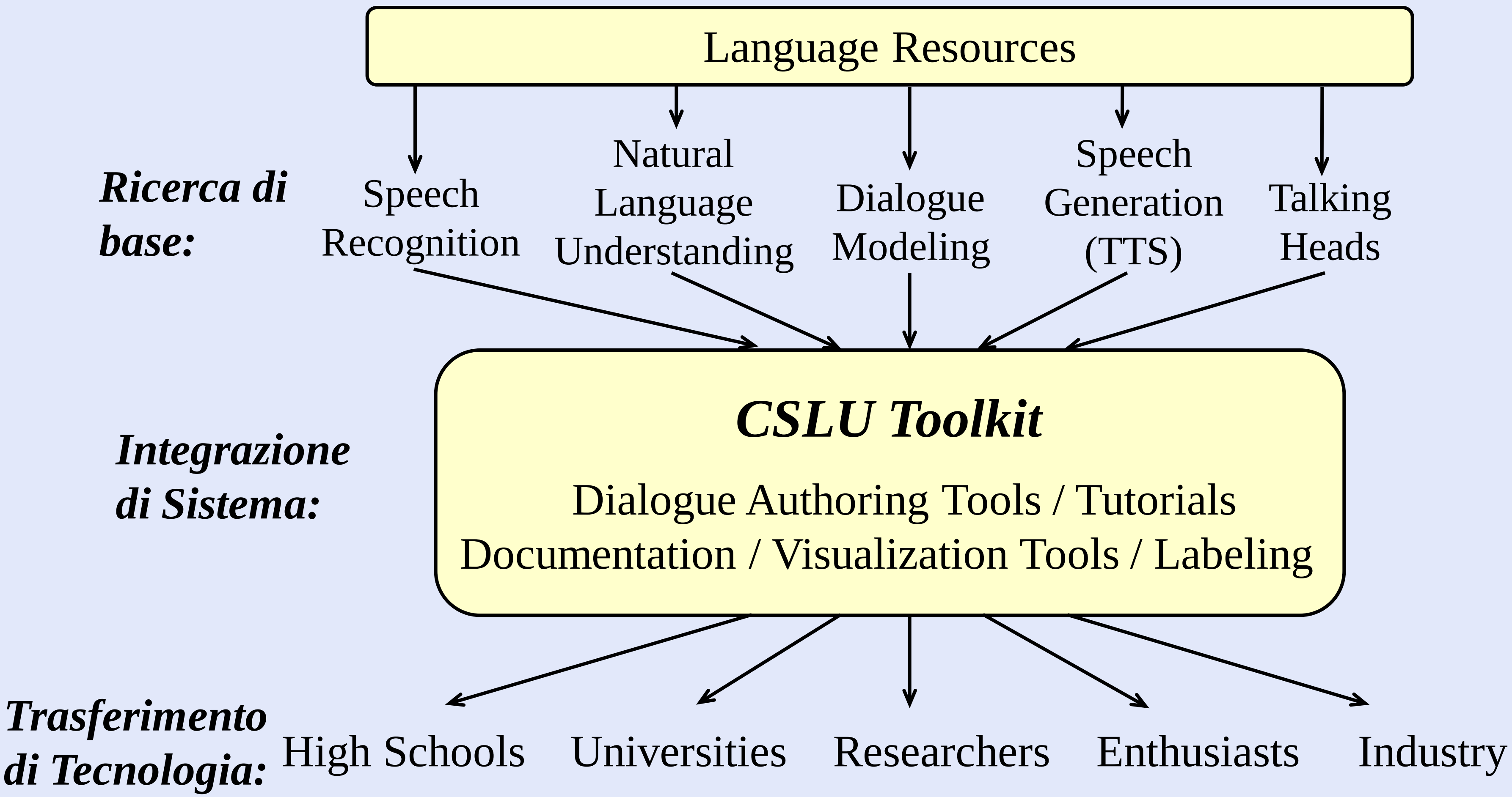
OREGON GRADUATE INSTITUTE OF SCIENCE AND TECHNOLOGY

CSLU Center for Spoken Language Understanding

John-Paul Hosom
P.O. Box 91000, Portland Oregon 97291-1000 USA
e-mail:hosom@cse.ogi.edu/

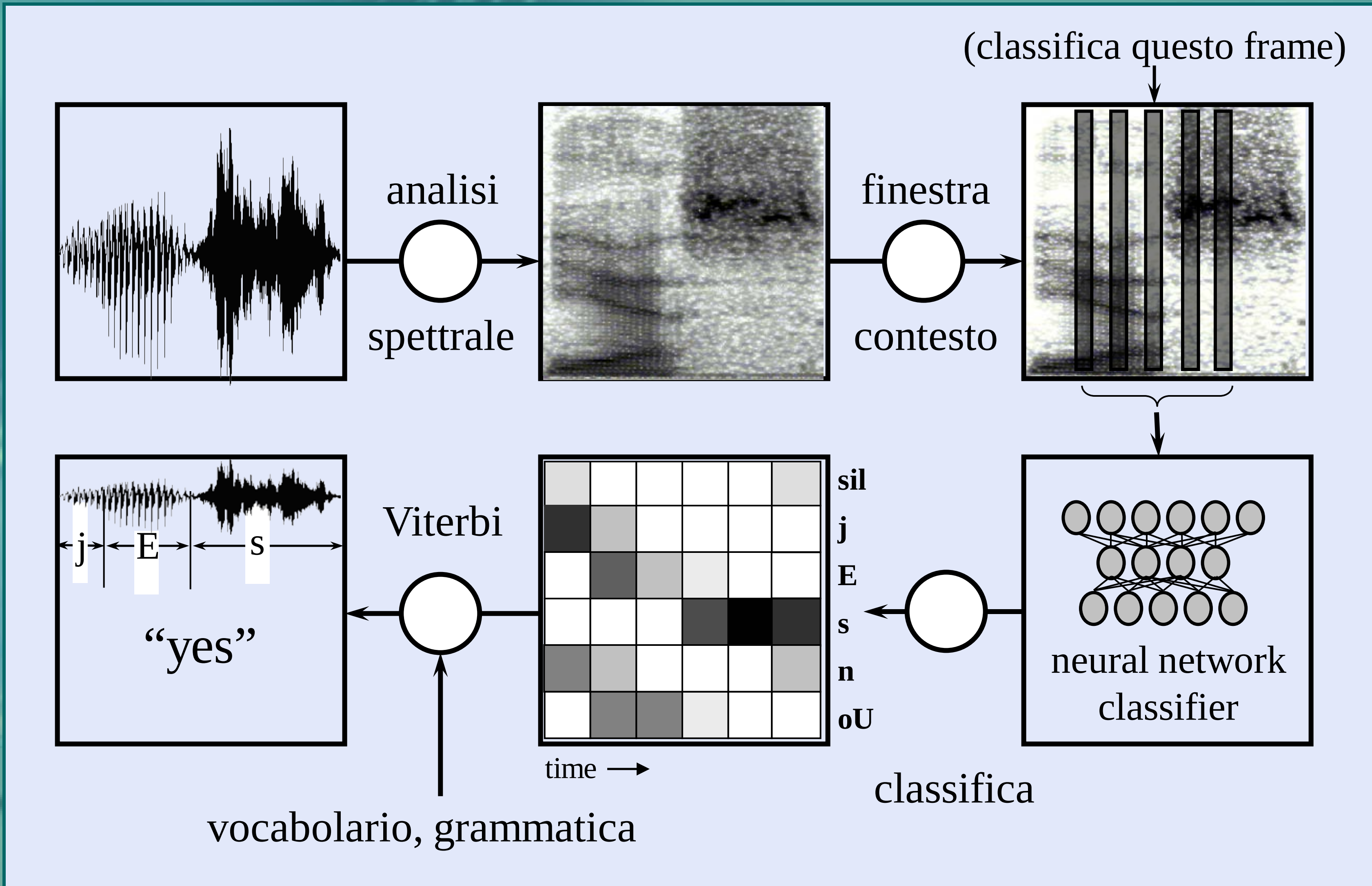
CLSU Toolkit: http://cslu.cse.ogi.edu/toolkit/

Cosa sono?



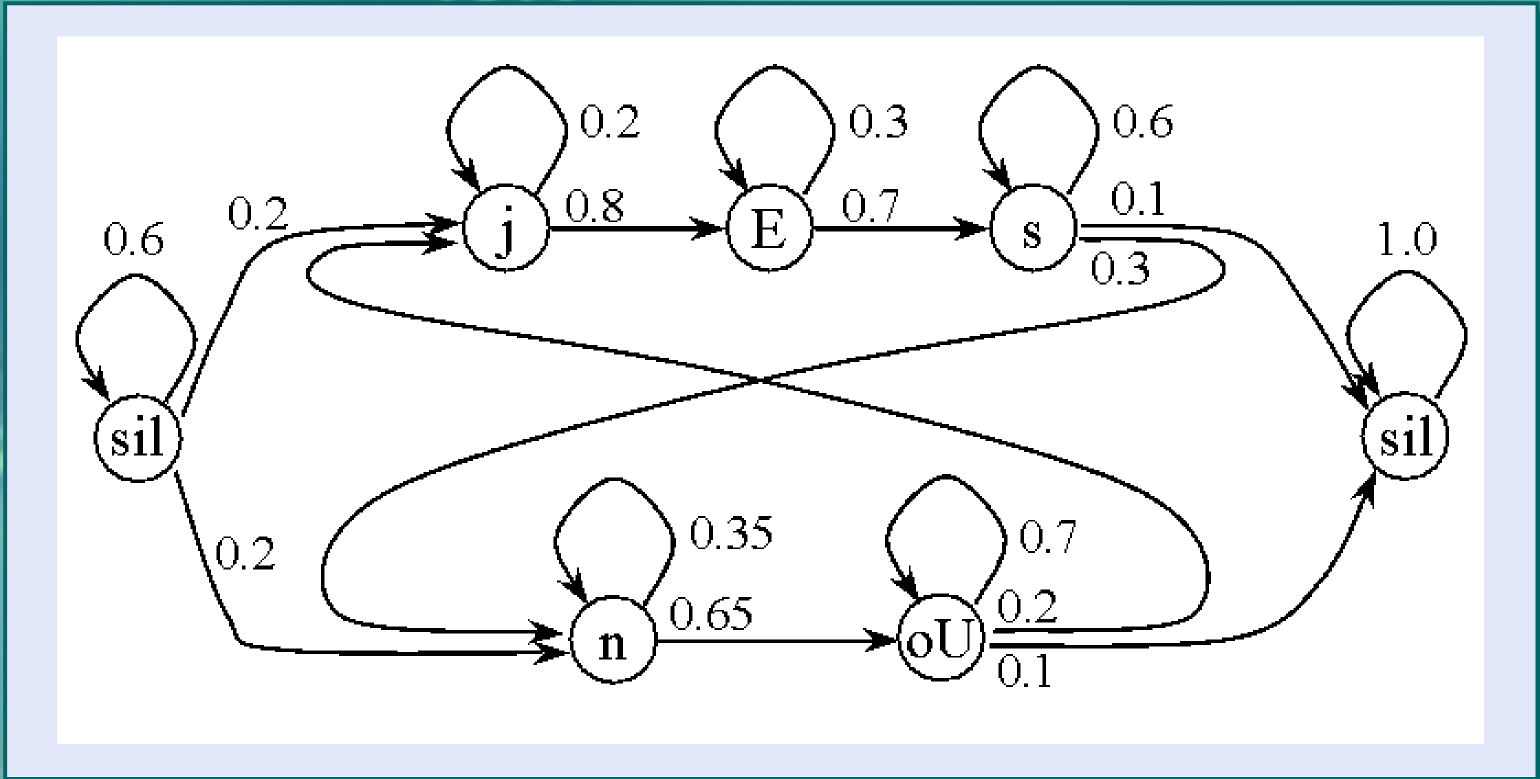
download: **http://cslu.cse.ogi.edu/toolkit/**

Il sistema di riconoscimento



Il sistema di riconoscimento

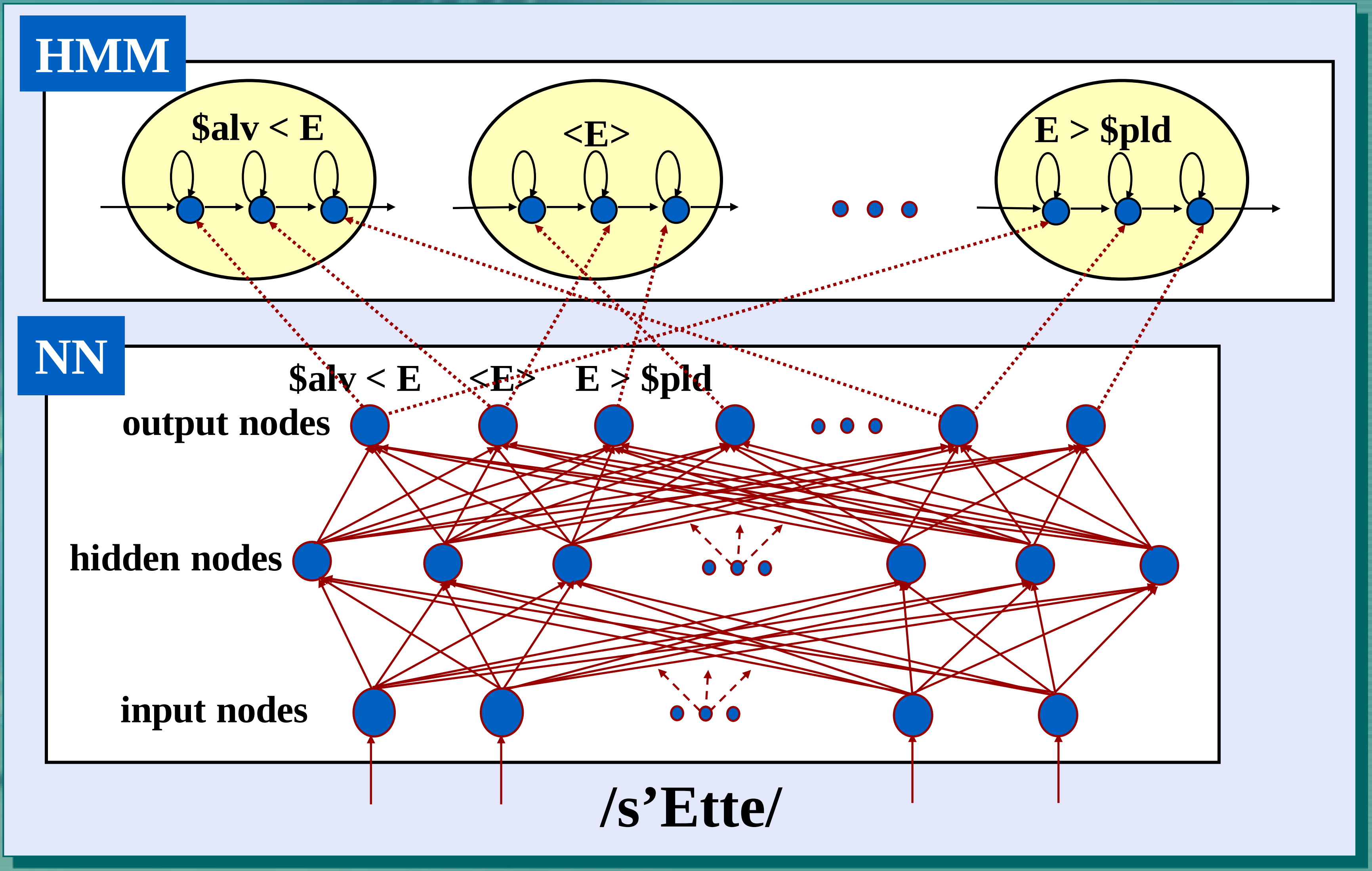
- **Standard HMM:**



Le probabilità degli stati sono stimate da ANN invece che da GMM

- **sistemi ibridi HMM/ANN (*cfr. Bourlard, Morgan*)**

Sistemi ibridi HMM/ANN



CSLU Speech Toolkit: EVALITA

- three-layer fully connected feed-forward network
- trained to estimate, at every frame, the probability of 98 context-dependent phonetic categories; these categories were created by splitting each Acoustic Unit (AU), into one, two, or three parts, depending on the length of the AU and how much the AU was thought to be influenced by co-articulatory effects. Silence and closure are 1-part units, vowels are 3-part units, unvoiced plosive is 1-part right dependent unit, voiced plosive, affricate, fricative, nasal, liquid retroflex and glide are all 2-part units
- 100 iterations; the best network iteration (baseline network - B) was determined by evaluation on the EVALITA clean and noisy digits development sets respectively
- after a comparison among the CSLU system driven by different feature types, we have found that 13-coefficient PLP plus 13-coefficient MFCC with CMS computed, once every 10 ms obtained the best score.

SONIC

University of Colorado, Center for Spoken Language Research

bpellom@rosettastone.com

pellom@cslr.colorado.edu

Boulder Language Technologies

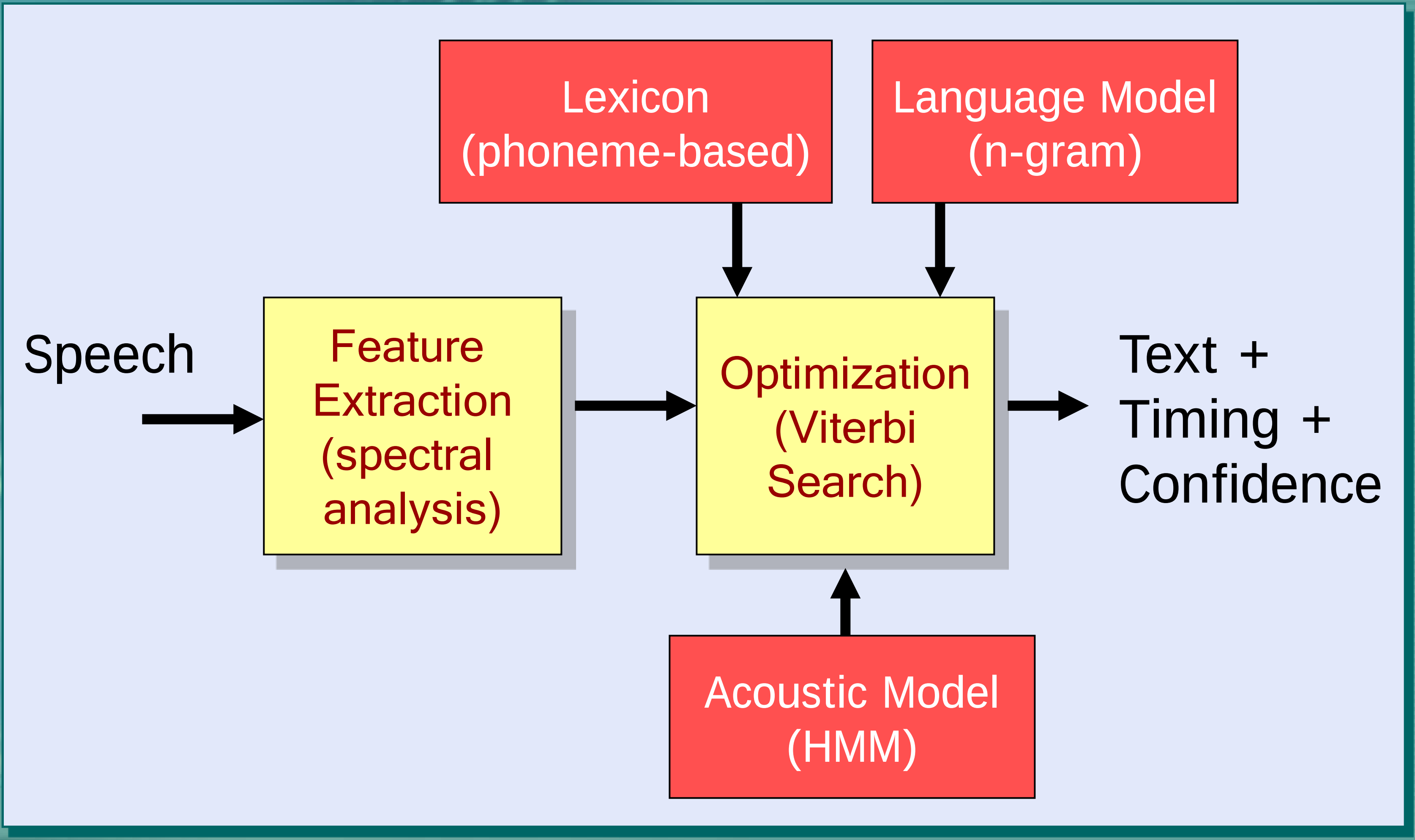
http://www.bltek.com/virtual-teacher-side-menu/sonic.html

ASR: sonic

- **Front-End**
- Lexicon
- **Acoustic Model**
- Language Model
- **Search**
- Adaptation

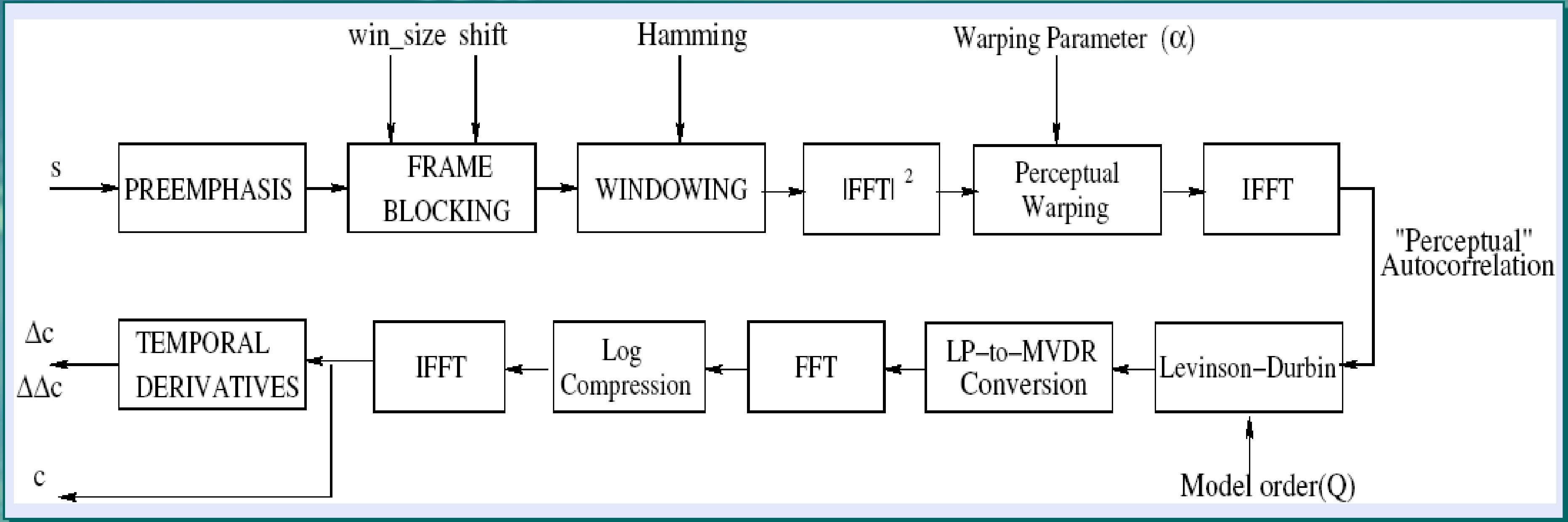
University of Colorado, Center for Spoken Language Research
(Bryan PELLOM)

ASR Structure



Feature Extraction: PMVDR

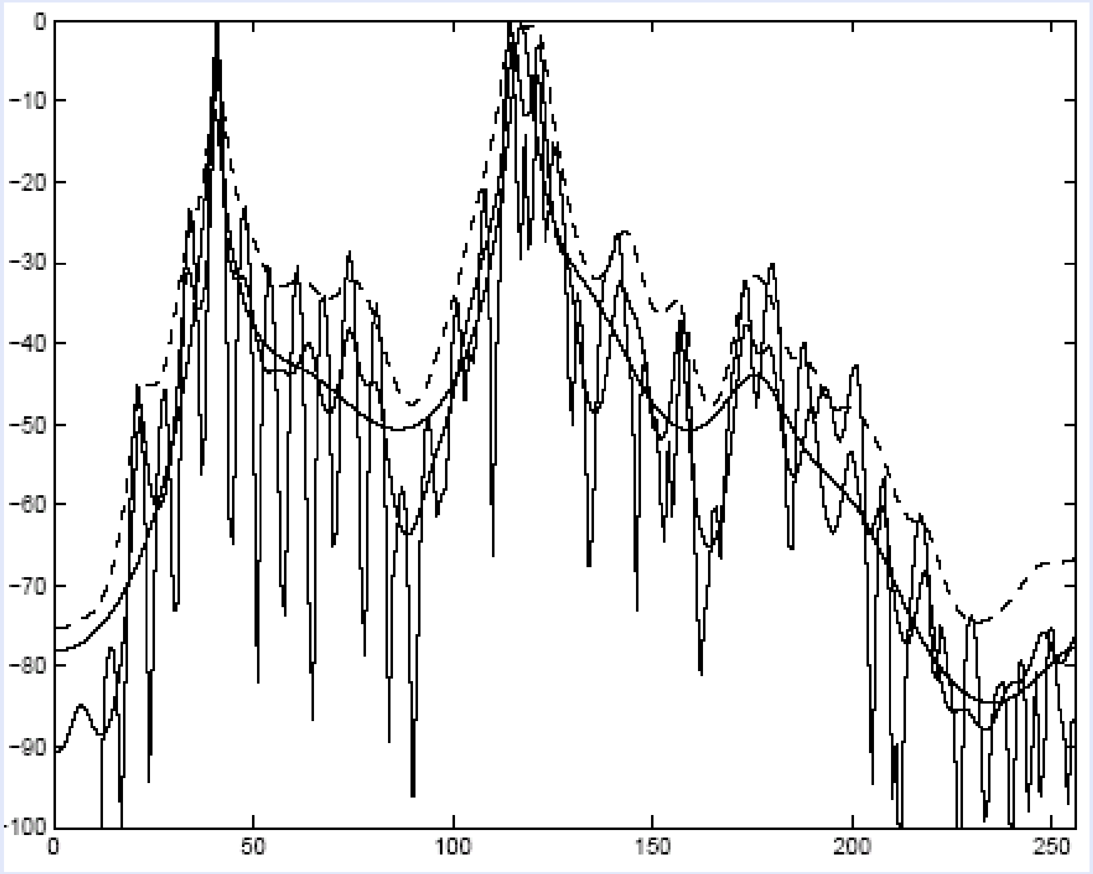
Perceptual Minimum Distortionless Response Cepstral Coefficients



(Yapanel & Hansen, Eurospeech 2003)

MVDR

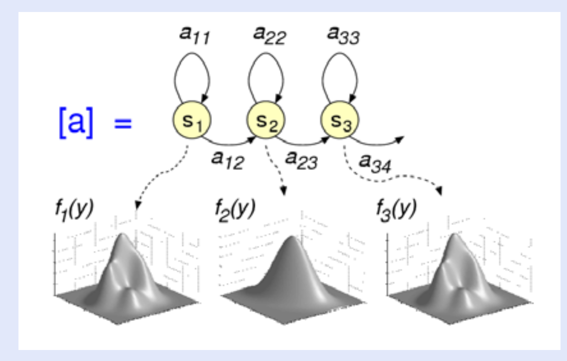
Minimum Variance Distortionless Response



Spectral envelopes: LP (solid), MVDR (dashed)

HMM

- the acoustic models consists of **decision-tree state-clustered HMMs** with associated **gamma probability density functions** to **model state-durations.**
- the acoustic models have a fixed **3-state topology**
- each HMM state can be modelled with variable number of multivariate mixture Gaussian distributions
- the training process consists of first performing state-based alignment of the training audio followed by an **expectation-maximization (EM) step** in which **decision-tree state-clustered HMMs are estimated**
- acoustic model parameters (means, covariances, and mixture weights) are estimated in the **maximum likelihood** sense
- the training process can be iterated between alignment of data and model estimation to gradually achieve adequate parameter estimation



Viterbi-based Training of Italian Speech Models

SAMPA	Example	SAMPA	Example	SAMPA	Example	SAMPA	Example
i	pini	i1	così	d	dente	dZ	magia
e	velo	e1	mercé	g	gatto	m	mano
E	aspetto	E1	caffè	f	faro	n	nave
a	vai	a1	bontà	s	sole	J	legna
o	polso	o1	Roma	S	sci	nf	anfora
O	cosa	O1	però	v	via	ng	ingordo
u	punta	U1	più	z	peso	l	palo
j	piume	t	torre	ts	pizza	L	soglia
w	quando	k	caldo	tS	pece	r	remo
p	pera	b	botte	dz	zero	SIL	silence

SONIC: EVALITA

- **12 PMVDR** cepstral parameters were retained and augmented with **normalized log frame energy** plus **Δ** and **ΔΔ**
- **39-dimensional feature vector** is computed, once every **10 ms**
- the model developed in [9] was inserted in the first alignment step to provide a good segmentation to start from and a first acoustic-model estimation was computed
- at the end of further **8 loops** of **phonetic alignment** and **acoustic model re-estimation**, the final AM is considered well trained

[9] Cosi, P., Hosom, J.P. (2000). High Performance "General Purpose" Phonetic Recognition for Italian, Proceedings of ICSLP 2000, International Conference on Spoken Language Processing, Beijing, China (2000), vol. II, pp. 527--530.

SPHINX

Carnegie Mellon University

School of Computer Science
Department of Electrical & Computer Engineering

http://www.cs.cmu.edu/

CMU SPHINX
http://cmusphinx.sourceforge.net/html/cmusphinx.php

SPHINX 3

- ***Lexical model***: The lexical or pronunciation model contains pronunciations for all the words of interest to the decoder. Sphinx-3 uses *phonetic units* to build word pronunciations. Currently, the pronunciation lexicon is almost entirely hand-crafted.
- ***Acoustic model***: Sphinx uses acoustic models based on statistical *hidden Markov models* (HMMs). The acoustic model is trained from acoustic training data using the Sphinx-3 trainer. The trainer is capable of building acoustic models with a wide range of structures, such as *discrete*, *semi-continuous*, or *continuous*. However, the s3.3 decoder is only capable of handling continuous acoustic models.
- ***Language model (LM)***: Sphinx-3 uses a conventional backoff bigram or trigram language model.

SPHINX 3

- training is an iterative sequence of alignments and AM-estimations; it starts from an audio segmentation aligned to training-data transcriptions and it estimates a raw first AM from them
- this is the starting point of the following loops of Baum-Welch probability density functions estimation and transcription alignment; models can be computed either for each phoneme (Contest Independent, CI) or, considering phoneme context (Contest Dependent, CD)
- SPHINX acoustic models are trained over MFCC + Δ + ΔΔ feature vectors
- SPHINX-3, is a C-based state-of-the–art large-vocabulary continuous-model ASR, and it is limited to 3 or 5-state left-to-right HMM topologies and to a bigram or trigram language model
- the decoder is based on the conventional Viterbi search algorithm and beam search heuristics
- it uses a lexical-tree search structure, too, in order to prune the state transitions

SPHINX 3: *EVALITA*

- no previously developed AM was applied and a simple uniform segmentation was chosen as starting point
- after a raw first-AM estimation, 4 loops of re-alignment and CI (contest-independent) AM re-estimations were done
- the last CI trained model was employed to create a minimum-error segmentation and train contest-dependent AMs
- an all-state (untied) AM was computed, and then 4 loops of CD state-tied segmentation–re-estimation were done

Results: CSLU Speech Toolkit

development WA %	IA	FA	FB1	FB2	FB3
clean AM on clean	99,82	99,75	**99,94**	99,82	99,75
noisy AM on noisy	90,15	90,93	**92,11**	91,75	91,49
full AM on clean+noisy	93,86	94,12	**94,28**	**94,28**	94,2

test FB1	WA %	SA %
clean AM on clean	**99,10**	**94,80**
noisy AM on noisy	**94,00**	**82,00**
full AM on clean + noisy	**95,00**	**87,20**

Results: SONIC

development WA %	full AM	clean AM	noisy AM
clean	99,70	**99,80**	**99,70**
noisy	94,20	89,90	**94,80**
clean + noisy	96,71	94,42	**97,04**

test	WA %	SA %
clean AM on clean	**99,60**	**97,30**
noisy AM on noisy	**96,30**	**87,90**
full AM on clean + noisy	**97,30**	**90,60**

Results: SPHINX 3

development WA %	full AM	clean AM	noisy AM
clean	**99,40**	**99,40**	98,80
noisy	**93,30**	78,70	92,60
clean + noisy	**96,10**	88,31	95,43

test	WA %	SA %
clean AM on clean	**98,90**	**94,50**
noisy AM on noisy	**91,70**	**72,70**
full AM on clean + noisy	**95,50**	**86,00**

Results

test	CSLR		SONIC		SPHINX	
	WA %	SA %	WA %	SA %	WA %	SA %
clean AM on clean	99,10	94,80	**99,60**	**97,30**	98,90	94,50
noisy AM on noisy	94,00	82,00	**96,30**	**87,90**	91,70	72,70
full AM on clean + noisy	95,00	87,20	**97,30**	**90,60**	95,50	86,00

Concluding Remarks

- **3 of the most used open source ASR tools were considered in this work, i.e. CSLU Toolkit, SONIC, and SPHINX, because promising results were obtained in the past on similar digit recognition tasks**
- **beyond the fact that finding similarity among the three ASR systems was one of the main difficulties, an homogeneous and unique test framework for comparing different Italian ASR systems was quite possible and effective if 3-gram LM weight is set to 0 and the results produced by the best WA-score configuration were compared for each system**
- **CSLU Toolkit is good in recognizing clean digit sequences, but it is not so good at recognizing clean-plus-noisy audio; SONIC is the best system in all situations and we believe this is mainly due to the adoption of the PMVDR features; SPHINX is quite more sensible to AM specialization than other systems and clean models can not recognize noisy speech with high performance.**
- **finally we should conclude that the EVALITA Speech campaign was quite effective in forcing various Italian research groups to focus on similar recognition tasks working on common data thus comparing and improving various different recognition methodologies and strategies**

Future Work

- **we hope more complex tasks and data will be exploited in the future**
- **we are looking for CHILDREN Speech evaluation campaign**
- **EVALEU**

  Interspeech 2011 - Satellite Workshop????

Thank You!!! …… and

WELCOME to Florence 2011