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

Copyright, 2009 © ISTC-SPFD-CNR
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
Data

<table>
<thead>
<tr>
<th>Sub Set</th>
<th>Clean Audio Files</th>
<th>Noisy Audio Files</th>
<th>Clean Digit Sequences</th>
<th>Noisy Digit Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>3144</td>
<td>2204</td>
<td>10129</td>
<td>7376</td>
</tr>
<tr>
<td>Development</td>
<td>216</td>
<td>299</td>
<td>1629</td>
<td>1941</td>
</tr>
<tr>
<td>Test</td>
<td>365</td>
<td>605</td>
<td>2361</td>
<td>4036</td>
</tr>
</tbody>
</table>

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.

simple grammar

[<any> (<digit> [silence]) + <any>]
Cosa sono?

**Ricerca di base:**
- Speech Recognition
- Natural Language Understanding
- Dialogue Modeling
- Speech Generation (TTS)
- Talking Heads

**Integrazione di Sistema:**
- CSLU Toolkit
  - Dialogue Authoring Tools / Tutorials
  - Documentation / Visualization Tools / Labeling

**Trasferimento di Tecnologia:**
- High Schools
- Universities
- Researchers
- Enthusiasts
- Industry

**Language Resources**

download: http://cslu.cse.ogi.edu/toolkit/
Il sistema di riconoscimento

analisi spettrale

finestra contesto

neural network classifier

classifica

glosario, grammatica

"yes"

Viterbi
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

HMM

NN

/output nodes

/hidden nodes

/input nodes

/s’Ette/
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.
University of Colorado, Center for Spoken Language Research

bpellom@rosettastone.com
pellom@cslr.colorado.edu

ASR: sonic

- Front-End
- Lexicon
- Acoustic Model
- Language Model
- Search
- Adaptation
ASR Structure

- **Speech**
- **Feature Extraction** (spectral analysis)
- **Optimization** (Viterbi Search)
- **Acoustic Model** (HMM)
- **Lexicon** (phoneme-based)
- **Language Model** (n-gram)

Output: Text + Timing + Confidence
Feature Extraction: PMVDR

Perceptual Minimum Distortionless Response

Cepstral Coefficients

(Yapanal & Hansen, Eurospeech 2003)
MVDR

Minimum Variance Distortionless Response

Spectral envelopes: LP (solid), MVDR (dashed)
The acoustic models consist 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 a 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

<table>
<thead>
<tr>
<th>SAMPA</th>
<th>Example</th>
<th>SAMPA</th>
<th>Example</th>
<th>SAMPA</th>
<th>Example</th>
<th>SAMPA</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>pini</td>
<td>il</td>
<td>cosi</td>
<td>d</td>
<td>dente</td>
<td>dZ</td>
<td>magia</td>
</tr>
<tr>
<td>e</td>
<td>velo</td>
<td>el</td>
<td>mercе</td>
<td>g</td>
<td>gatto</td>
<td>m</td>
<td>mano</td>
</tr>
<tr>
<td>E</td>
<td>aspetto</td>
<td>El</td>
<td>caffе</td>
<td>f</td>
<td>faro</td>
<td>n</td>
<td>nave</td>
</tr>
<tr>
<td>a</td>
<td>vai</td>
<td>al</td>
<td>bonta</td>
<td>s</td>
<td>sole</td>
<td>J</td>
<td>legna</td>
</tr>
<tr>
<td>o</td>
<td>polso</td>
<td>ol</td>
<td>Roma</td>
<td>S</td>
<td>sci</td>
<td>nf</td>
<td>anforа</td>
</tr>
<tr>
<td>O</td>
<td>cosa</td>
<td>Ol</td>
<td>perо</td>
<td>v</td>
<td>via</td>
<td>ng</td>
<td>ingordo</td>
</tr>
<tr>
<td>u</td>
<td>punta</td>
<td>Ul</td>
<td>più</td>
<td>z</td>
<td>peso</td>
<td>l</td>
<td>palо</td>
</tr>
<tr>
<td>j</td>
<td>piumе</td>
<td>t</td>
<td>torre</td>
<td>ts</td>
<td>pizza</td>
<td>L</td>
<td>soglia</td>
</tr>
<tr>
<td>w</td>
<td>quando</td>
<td>k</td>
<td>caldo</td>
<td>tS</td>
<td>pece</td>
<td>r</td>
<td>remo</td>
</tr>
<tr>
<td>p</td>
<td>perа</td>
<td>b</td>
<td>botte</td>
<td>dz</td>
<td>zero</td>
<td>SIL</td>
<td>silence</td>
</tr>
</tbody>
</table>
12 PMVDR cepstral parameters were retained and augmented with normalized log frame energy plus $\Delta$ and $\Delta\Delta$

A 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.

Carnegie Mellon University

School of Computer Science
Department of Electrical & Computer Engineering

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

CMU SPHINX
**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.
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.
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**

<table>
<thead>
<tr>
<th>development WA %</th>
<th>IA</th>
<th>FA</th>
<th>FB1</th>
<th>FB2</th>
<th>FB3</th>
</tr>
</thead>
<tbody>
<tr>
<td>clean AM on clean</td>
<td>99,82</td>
<td>99,75</td>
<td>99,94</td>
<td>99,82</td>
<td>99,75</td>
</tr>
<tr>
<td>noisy AM on noisy</td>
<td>90,15</td>
<td>90,93</td>
<td>92,11</td>
<td>91,75</td>
<td>91,49</td>
</tr>
<tr>
<td>full AM on clean+noisy</td>
<td>93,86</td>
<td>94,12</td>
<td>94,28</td>
<td>94,28</td>
<td>94,2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>test FB1</th>
<th>WA %</th>
<th>SA %</th>
</tr>
</thead>
<tbody>
<tr>
<td>clean AM on clean</td>
<td>99,10</td>
<td>94,80</td>
</tr>
<tr>
<td>noisy AM on noisy</td>
<td>94,00</td>
<td>82,00</td>
</tr>
<tr>
<td>full AM on clean + noisy</td>
<td>95,00</td>
<td>87,20</td>
</tr>
</tbody>
</table>
## Results: SONIC

<table>
<thead>
<tr>
<th>development WA %</th>
<th>full AM</th>
<th>clean AM</th>
<th>noisy AM</th>
</tr>
</thead>
<tbody>
<tr>
<td>clean</td>
<td>99,70</td>
<td>99,80</td>
<td>99,70</td>
</tr>
<tr>
<td>noisy</td>
<td>94,20</td>
<td>89,90</td>
<td>94,80</td>
</tr>
<tr>
<td>clean + noisy</td>
<td>96,71</td>
<td>94,42</td>
<td>97,04</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>test</th>
<th>WA %</th>
<th>SA %</th>
</tr>
</thead>
<tbody>
<tr>
<td>clean AM on clean</td>
<td>99,60</td>
<td>97,30</td>
</tr>
<tr>
<td>noisy AM on noisy</td>
<td>96,30</td>
<td>87,90</td>
</tr>
<tr>
<td>full AM on clean + noisy</td>
<td>97,30</td>
<td>90,60</td>
</tr>
</tbody>
</table>
### Results: SPHINX 3

#### Development WA %

<table>
<thead>
<tr>
<th></th>
<th>full AM</th>
<th>clean AM</th>
<th>noisy AM</th>
</tr>
</thead>
<tbody>
<tr>
<td>clean</td>
<td>99,40</td>
<td>99,40</td>
<td>98,80</td>
</tr>
<tr>
<td>noisy</td>
<td>93,30</td>
<td>78,70</td>
<td>92,60</td>
</tr>
<tr>
<td>clean + noisy</td>
<td>96,10</td>
<td>88,31</td>
<td>95,43</td>
</tr>
</tbody>
</table>

#### Test WA %

<table>
<thead>
<tr>
<th></th>
<th>WA %</th>
<th>SA %</th>
</tr>
</thead>
<tbody>
<tr>
<td>clean AM on clean</td>
<td>98,90</td>
<td>94,50</td>
</tr>
<tr>
<td>noisy AM on noisy</td>
<td>91,70</td>
<td>72,70</td>
</tr>
<tr>
<td>full AM on clean + noisy</td>
<td>95,50</td>
<td>86,00</td>
</tr>
<tr>
<td>test</td>
<td>CSLR</td>
<td>SONIC</td>
</tr>
<tr>
<td>------------------------</td>
<td>------</td>
<td>-------</td>
</tr>
<tr>
<td></td>
<td>WA %</td>
<td>SA %</td>
</tr>
<tr>
<td>clean AM on clean</td>
<td>99,10</td>
<td>94,80</td>
</tr>
<tr>
<td>noisy AM on noisy</td>
<td>94,00</td>
<td>82,00</td>
</tr>
<tr>
<td>full AM on clean + noisy</td>
<td>95,00</td>
<td>87,20</td>
</tr>
</tbody>
</table>
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