EVALITA 2009

Connected Digits Recognition Task

Gianpaolo Coro*, Roberto Gretter°, Marco Matassoni°

* ABLA

° FBK FONDAZIONE BRUNO KESSLER

Task Description

In the Connected Digits Recognition Task, systems are required to recognize digits sequences uttered in a speech signal.

The task consists of two tracks:

- *Clean Speech Digit Sequence Recognition Task*: recognize digits sequences in **clean** speech environment.

- *Noisy Speech Digit Sequence Recognition Task:* recognize digits sequences in **noisy** speech environment.
Noise may vary from white noise to traffic, room, etc.

Task Motivation

Compare recognition systems focusing only on some components:

- small active dictionary:
  - reduce development time and training data collection and distribution.
  - almost independent from language model

- problems that can be found in more complex tasks
  - continuous speech
  - shared phonemes across words

Connected digits sequences automatic recognition focus more on acoustic models and feature representation, neglecting language model.

Corpus Description 1/2

The corpus has been taken from various Italian acoustic corpora.

- Speakers are almost equally distributed along the territory
- Annotation at sentence level is provided
- Audio files are sampled at 16 kHz, 16 bit PCM, mono and stored in Windows .wav format
- Release of training, development and test sets

Corpus Description 2/2

Clean Sets	Sentences	Speakers	# digits	Length
Train	3144	300	**10129**	~2h40m
Development	216	85	1629	~18m
Test	365	85	2360	~28m

Noisy Sets	Sentences	Speakers	# digits	Length
Train	2204	310	**7376**	~2h17m
Development	299	110	1940	~25m
Test	605	110	4036	~52m

Evaluation

Word Accuracy is defined as

$$WA = 100 - \frac{I + S + D}{N} \times 100$$

where, referring to the automatic transcription:
- I is the number of inserted words
- S is the number of substitutions
- D is the number of the deletions
- N is the number of words in the reference

Sentence Accuracy: is defined as

$$SA = \frac{H}{M} \times 100$$

Again, referring to the automatic transcription:
- H is the number of sentences correctly recognized
- M is the number of sentences in the reference

The evaluation is based on Minimum Edit Distance calculation between the transcription coming out from the recognizer and the orthographic annotation.

Participants

ABLA srl

CEDAT85

Istituto di Scienze e Tecnologie della Cognizione (ISTC-CNR)

University Federico II of Naples (only for Clean Speech Task)

Results in Clean Environment

Sentence Acc %	Word Acc %	Words	Del+ Ins+ Sub	System		Description
96.44	**99.45**	2360	7+6+0	**ISTC-SONIC_2**		• HMM Acoustic Models • Phonetic Approach • PMVDR Features • Decision-Tree State-Clustered HMMs • Trained on Clean Data
96.44	**99.45**	2360	8+3+2	**ISTC-SONIC_1**		• Structure as in ISTC-SONIC_2 • Trained on all the training data (Noisy + Clean)
96.16	**99.32**	2360	4+8+4	**ISTC-SPHINX_1**		• HMM Acoustic Models • Phonetic Approach • MFCC Features • Lexical Tree Search Structure • Trained on all the training data (Noisy + Clean)
95.89	**99.28**	2360	6+2+9	**ABLA-NUANCE**	T	• HMM Acoustic Models • Phonetic Approach • MFCC Features • Word Graphs Decoding • Big Training Data Set
95.62	**99.19**	2360	6+5+8	**ISTC-CSLU_1**		• HMM + ANN Acoustic Models • Phonetic Approach • MFCC+PLP Features • Trained on all the training data (Noisy + Clean)
94.25	**98.94**	2360	11+7+7	**ISTC-CSLU_2**		• Structure as in ISTC-CSLU_1 • Trained on Clean Data
93.70	**98.77**	2360	6+14+9	**ISTC-SPHINX_2**		• Structure as in ISTC-SPHINX_1 • Trained on Clean Data
89.59	**98.05**	2360	5+19+22	**CEDAT85 (Based on IBM VoiceTaylor)**	T	• HMM Acoustic Models • Phonetic Approach • Big Training Data Set in Clean Env.
81.64	**96.06**	2360	34+3+56	**ABLA-TSPEECH**	T	• HMM Acoustic Models • Syllabic Dynamic Approach • Energy and Duration Templates • Small Training Data Set (2000 words) in Clean Env.
18.36	**77.84**	2360	116+104+303	**UNINA**	L	• SVM Unity Classification • Automatic Syllabic Segmentation • Unit Graph Decoding • Trained on Clean Data

Results in Noisy Environment

Sentence Acc %	Word Acc %	Words	Del+ Ins+ Sub	System		Description
87.77	**96.21**	4036	104+13+36	**ISTC-SONIC_2**		• **HMM Acoustic Models** • **Phonetic Approach** • **PMVDR Features** • **Decision-Tree State-Clustered HMMs** • **Trained on Noisy Data**
86.45	**95.91**	4036	105+11+49	**ISTC-SONIC_1**		• **Structure as in ISTC-SONIC_2** • **Trained on all the training data (Noisy + Clean)**
81.82	**93.95**	4036	121+29+94	**ISTC-CSLU_2**		• **HMM + ANN Acoustic Models** • **Phonetic Approach** • **MFCC+PLP Features** • **Trained on Noisy Data**
79.17	**93.06**	4036	136+51+93	**ISTC-SPHINX_1**		• **HMM Acoustic Models** • **Phonetic Approach** • **MFCC Features** • **Lexical Tree Search Structure** • **Trained on all the training data (Noisy + Clean)**
81.65	**92.42**	4036	135+37+134	**ISTC-CSLU_1**		• **Structure as in ISTC-CSLU_2** • **Trained on all the training data (Noisy + Clean)**
72.56	**91.63**	4036	133+81+124	**ISTC-SPHINX_2**		• **Structure as in ISTC-SPHINX_1** • **Trained on Noisy Data**
78.02	**91.03**	4036	255+36+71	**CEDAT85 (Based on IBM VoiceTaylor)**	**T**	• **HMM Acoustic Models** • **Phonetic Approach** • **Big Training Data Set in Clean Env.**
77.69	**88.65**	4036	268+26+164	**ABLA-NUANCE**	**T**	• **HMM Acoustic Models** • **Phonetic Approach** • **MFCC Features** • **Word Graphs Decoding** • **Big Training Data Set**
69.09	**82.23**	4036	467+56+194	**ABLA-TSPEECH**	**T**	• **HMM Acoustic Models** • **Syllabic Dynamic Approach** • **Energy and Duration Templates** • **Small Training Data Set (2000 words) in Clean Env.**

Conclusion

Little mismatch between training and test data:

- very effective acoustic model

Open discussion about the effectiveness of various approaches to speech recognition:

- syllabic versus phonetic modelling
- choice of suitable acoustic features