EVALITA 2009: Description and Results of the Speech Recognition task

Gianpaolo Coro[1], Roberto Gretter[2], and Marco Matassoni[2]

[1]Abla srl, Viale Fulvio Testi, 7 Milano 20159
[2]FBK-irst, Povo via Sommarive 18, 38100 - Trento
gianpaolo.coro@abla.it, {gretter,matasso}@fbk.eu

Abstract. In this paper, we describe motivations and features of the Speech Recognition task at EVALITA 2009. Systems are compared about the performance on the recognition of uttered connected digits sequences. Interesting results will be shown for various types of approaches, ranging from *classic* algorithms, to non-standard models. Commercial as well as prototype speech recognizers have been involved into this task and results give an overview of the performance, which leave space to a more general discussion about approaches to speech recognition tasks.

Keywords: Speech recognition, Italian connected digits recognition, EVALITA 2009.

1 Introduction

This report presents the first Speech Recognition task organized for the EVALITA campaigns. The task was introduced to make a comparison among different Automatic Speech Recognizers (ASR), based on several technological approaches. Commercial, open source or prototype ASRs have been tested on a single task, the uttered digits sequences recognition [1, 2].

There are many motivations for such a task. In experiments about speech recognition it is often necessary to define an active dictionary, which has not to be too large, because it would cost too much in terms of development time and training data collection and distribution. In order to focus only on acoustic modeling, task with no language model can be conceived: the domain of (Italian) connected digits is interesting because some phonemes are shared among words, and connected speech problems have to be managed.

This popular task is well suited to compare novel approaches against standard algorithms. Moreover, in many commercial applications, in particular for human-machine interaction, users are frequently asked to utter a digits sequence, so that this experiment can be interesting even from an applicative perspective [3, 4].

The paper is organized as follows. Section 2 presents the definition of the task and the corpus, while Sections 3 and 4 describe evaluation measures and results. Finally, some conclusions are drawn.

2 Task Definition

In the Speech Recognition task, systems are required to recognize digits sequences (from 0 to 9) uttered in a speech signal. The task consists of two subtasks:

- in the *Clean Speech Digit Sequence Recognition Task*, systems are required to recognize digits sequences in clean speech environment.
- in the *Noisy Speech Digit Sequence Recognition Task*, systems are required to recognize digits sequences in noisy speech environment.

In the latter case, the type of noise may vary from white noise to traffic, room, etc. The selected corpus has been chosen to be very challenging, as the SNR can vary in type and in amplitude across the training and test files.

Some participants have trained their system on the provided training set, others have chosen to use other corpora for training, or to use pre-existing trained acoustic models, e.g., using commercial systems. In the latter case, two factors can affect accuracy: the amount of training data, that can be much larger, and the possible mismatch between recording conditions, especially for the noisy track.

The task is interesting because it allows to make a comparison among well-established approaches to speech recognition and non-standard frameworks.

Interesting considerations can be made on systems which used a single training corpus and challenged also on noisy environment.

2.1 Corpus Description

The corpus has been taken from various Italian acoustic corpora. Speakers are almost equally distributed along the territory; in particular a subdivision of Italy in zones has been made. The following is a schema of such classification:

- North: Valle d'Aosta, Piemonte, Lombardia, Liguria.
- North-East: Veneto, Trentino Alto Adige, Friuli Venezia Giulia.
- Center: Emilia Romagna, Toscana, Marche.
- Center-South: Lazio, Molise, Umbria, Abruzzo.
- South: Campania, Basilicata, Puglia, Calabria.
- Islands: Sardegna, Sicilia.

Annotation at sentence level is provided both for isolated and connected digits, and recordings come from telephone numbers and digits sequences uttered at microphone.

Audio files are sampled at 16 kHz, 16 bit PCM, mono and stored in Windows .wav format.

The clean environment consists of a large variety of speakers and signals: the training set contains 300 speakers, who uttered isolated as well as connected sequences. The development and test sets are based upon 85 speakers each. The localization of the speakers has been selected among the above described Italian zones, keeping a uniform geographic distribution.

The noisy environment contains about 310 speakers for the training set, and 110 for the development and test sets. Even in this case the localization was uniformly distributed among the Italian zones.

For the training and development sets, transcription at word level has been provided in two separate text files, containing on each row the audio filename followed by the transcription. Note that time segmentation was not available.

E.g. training.txt:

```
clean00001.wav 1 3 5 6 4 0
clean00002.wav 3 5 7 9
..
noisy02928.wav 2 4 6 8 3
```

Test data have been provided in the form of audio files and two lists of filenames (clean and noisy).

E.g. test_clean.txt:

```
clean00201.wav
clean00202.wav
..
noisy03109.wav
```

Tables 1 and 2 summarize the details of the proposed datasets for the clean and the noisy tasks respectively. Figures 1 and 2 show a couple of waveforms, along with the corresponding spectrograms, belonging to the two subtasks.

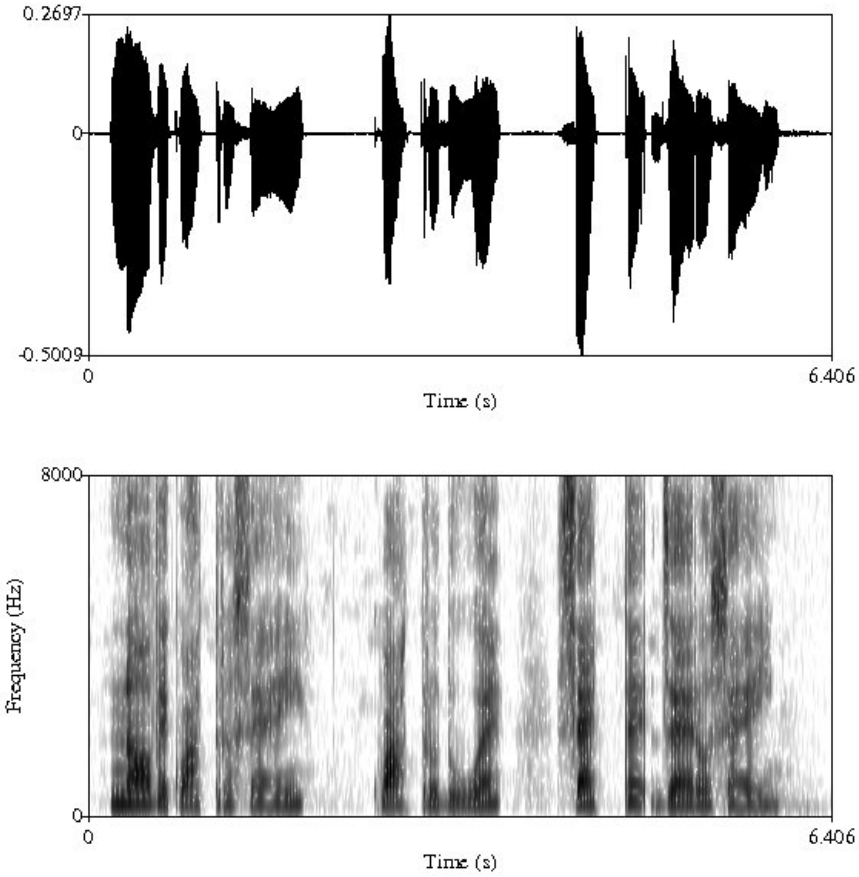

Fig. 1. Example of a clean digit sequence and the related spectrogram.

Table 1. Description of the datasets adopted in the EVALITA *clean* connected digit task.

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

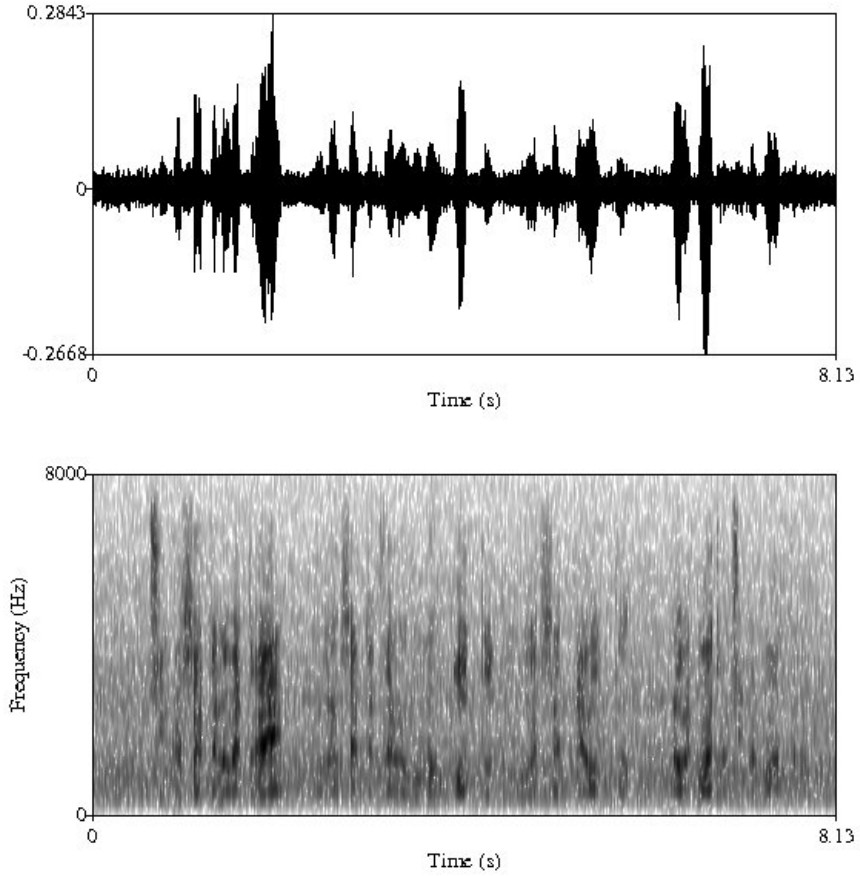

Fig. 2. Example of a noisy digit sequence and the related spectrogram.

Table 2. Description of the datasets adopted in the EVALITA *noisy* connected digit task.

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

3 Evaluation Measures

With respect to the results submitted by the participants, measurements of Word Accuracy and Sentence Accuracy are provided.

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.

Results for the two subtasks are provided in Tables 3 and 4. Some of the submissions, which appear in the Tables, should not be considered in the official results, either because training material different from the one provided was used (T, mnemonic for training) or because the submission was done after the deadline (L, mnemonic for Late).

The participants to this context were 4: UNINA, CEDAT85, ISTC, ABLA. All of them submitted results for the two subtasks, apart UNINA which participated only to the clean track. In total we had ten submissions for the clean track and nine for the noisy one.

Table 3. Results for the **Clean ASR task**, ordered by Word Accuracy. In the last column, T means that a non-official training was used, L means that the results were delivered late.

Sentence Acc	Word Acc	Words	Corr	Err	Del+Ins+Sub	System	
96.44%	**99.45%**	2360	2353	13	7+6+0	ISTC-SONIC_2	
96.44%	**99.45%**	2360	2350	13	8+3+2	ISTC-SONIC_1	
96.16%	**99.32%**	2360	2352	16	4+8+4	ISTC-SPHINX_1	
95.89%	**99.28%**	2360	2345	17	6+2+9	ABLA-NUANCE	**T**
95.62%	**99.19%**	2360	2346	19	6+5+8	ISTC-OGI_1	
94.25%	**98.94%**	2360	2342	25	11+7+7	ISTC-OGI_2	
93.70%	**98.77%**	2360	2345	29	6+14+9	ISTC-SPHINX_2	
89.59%	**98.05%**	2360	2333	46	5+19+22	CEDAT85	**T**
81.64%	**96.06%**	2360	2270	93	34+3+56	ABLA-TSPEECH	**T**
18.36%	**77.84%**	2360	1941	523	116+104+303	UNINA	**L**

Table 4. Results for the **Noisy ASR task**, ordered by Word Accuracy. In the last column, **T** means that a non-official training was used.

Sentence Acc	Word Acc	Words	Corr	Err	Del+Ins+Sub	System	
87.77%	**96.21%**	4036	3896	153	104+13+36	ISTC-SONIC_2	
86.45%	**95.91%**	4036	3882	165	105+11+49	ISTC-SONIC_1	
81.82%	**93.95%**	4036	3821	244	121+29+94	ISTC-OGI_2	
79.17%	**93.06%**	4036	3807	280	136+51+93	ISTC-SPHINX_1	
81.65%	**92.42%**	4036	3767	306	135+37+134	ISTC-OGI_1	
72.56%	**91.63%**	4036	3779	338	133+81+124	ISTC-SPHINX_2	
78.02%	**91.03%**	4036	3710	362	255+36+71	CEDAT85	**T**
77.69%	**88.65%**	4036	3604	458	268+26+164	ABLA-NUANCE	**T**
69.09%	**82.23%**	4036	3375	717	467+56+194	ABLA-TSPEECH	**T**

As the organizers don't know the details about the recognition engines, it is quite difficult to comment the tables and we can only make the following observations:

- System UNINA is based on a novel and alternative paradigm and this can partly explain the huge gap with the other systems;
- Systems trained on different corpora did not show top performance; it demonstrates that the provided training data are sufficient for build an effective system.

4 Discussion

In the EVALITA 2009 Speech Recognition Task systems are compared on performance related to recognition of uttered connected digits sequences.

The main evidence from the results is the strong impact of the type of training material on performance: the systems that used the provided training sets generally performed better that the others. Since the task was conceived to present little mismatch between training and test data, the exploitation of the whole training set allows to come up with a very effective acoustic model.

Moreover in noisy conditions the performance of the best systems is satisfactory, comparable to the very high results obtained in clean conditions.

References

1. Leonard, R. G.,: A Database for Speaker-Independent Digit Recognition. In: Proceedings of ICASSP, vol. 3, pp. 42.11 (1984)
2. Falavigna D., Gretter R.: On Field Experiments of Continuous Digit Recognition over the Telephone Network. In: Proceedings of EUROSPEECH. Rhodes, Greece (1997)
3. Rahim M.: Recognizing connected digits in a natural spoken dialog. In: Proceedings of ICASSP, vol. 1, pp. 153--156 (1999)
4. Cosi, P., Hosom, J. P., Tesser, F.: High Performance Italian Continuous Digit Recognition. In: Proceedings of International Conference on Spoken Language Processing, vol. IV, pp. 242--245. Beijing, China (2000)