

# EVALITA 2009: Abla srl Participant Report

Leandro D'Anna<sup>1</sup>, Gianpaolo Coro<sup>1</sup>, and Francesco Cutugno<sup>2</sup>

<sup>1</sup> Abla srl, Viale Fulvio Testi, 7 Milano 20159

<sup>2</sup> Università Federico II di Napoli

ldanna@unisa.it, gianpaolo.coro@abla.it, cutugno@unina.it

**Abstract.** In this paper we describe the two systems we presented at the EVALITA 2009 workshop, for the connected digits recognition task. The former is an Abla srl proprietary speech recognizer, based on standard decoding algorithms, with syllabic acoustic models. The recognition phase is followed by a rescoring session, based on syllables energy and duration templates, which recover some recognition errors. The latter recognizer is a system, based on standard algorithms and triphonic acoustic models. The core comes from Nuance ASR 8.5 recognizer, trained on a proprietary corpus. Performances are discussed and compared to the best performing EVALITA 2009 system on this task.

**Keywords:** Speech recognition, ASR, Abla, TSpeech, Nuance, connected digits recognition, EVALITA 2009, evaluation of NLP systems.

## 1 Introduction

In many applications, especially in telephony environment, users deal with automatic responders (IVR) with simple steps of interactions. Not often a natural language speech recognition is required, because people are not used to such technology, especially when the service addresses to a large group of persons.

Abla srl wanted to build up a low cost speech recognizer, tied on the most common usage of speech recognition, which did not require statistical language models, or a huge quantity of phonemes combinations to be trained.

Syllables were chosen for this task. A subset of recognition tasks (e.g. digits, commands for mail reading etc.) was selected, and a speech recognizer was born. Such product has been called TSpeech [1], and the digits recognition task of EVALITA 2009, was suited to test its effectiveness.

Syllables are robust to noise and speakers variations [2] and then it has been interesting to test how they performed on a corpus on which they had partially been trained. A further module, based on syllables energy and duration, was added to manage some recognition errors, in order to rise up the syllabic recognizer performances.

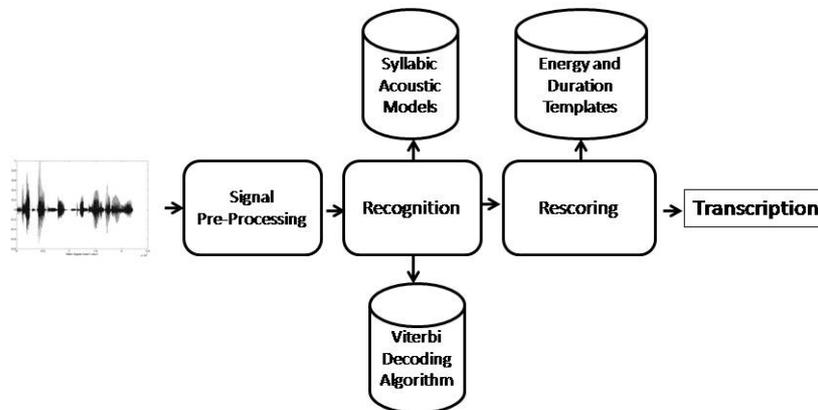
A second approach will be also presented here. The Nuance 8.5 ASR [3] was taken as the core model for another system we participated with. We chose to calculate performances with this system for two reasons. Abla is a commercial partner of

Nuance, so that we chose to participate together, as this solution has been used in many projects we managed. Moreover, we considered useful to show the performances of a state of the art commercial product on this task, in order to give TSpeech a sort of baseline.

The paper is organized as follows. Section 2 presents the systems description, while Sections 3 shows the results on the EVALITA 2009 test corpus, and in Section 4 some conclusions are drawn.

## 2 System Description

In this section we describe the architecture of the systems we used for the workshop.



**Fig. 1.** TSpeech architecture.

Figure 1, depicts the architecture of TSpeech. The speech signal passes through a signal preprocessing phase, which uses a voice activity detector, based on energy, to cut the audio waveform. The preprocessing also takes care of bursts and signal saturation. The formers are deleted, while for the latter a de-emphasis filter is applied.

The recognition module performs a transcription of the speech signal. This phase uses syllabic acoustic models, trained on a corpus of about 200 examples for each unit, to get a transcription with markers indicating syllables boundaries.

Table 1 shows the digits syllables set we used. They were obtained from considerations about the acoustically desired energy shape of syllables, in which an onset, a nucleus and a coda have to be present. The resulting segmentation in syllables is then slightly different from the linguistic separation.

The Viterbi algorithm [4] is applied, in combination with a static language model, in order to get syllables sequence transcription.

**Table 1.** Syllable segmentation of digits for TSpeech acoustic models.

Syllable	Words sharing the syllable
u	uno
no	uno, nove
due	due
tre	tre
kwa	quattro
ttro	quattro
tSin	cinque
kwe	cinque
sei	sei
se	sette
tte	sette
o	otto
ve	nove

Syllabic acoustic models were trained on hand labeled corpus of Italian language, in collaboration with the Federico II University of Naples. The annotation was made at syllable level, indicating noise and silence zones.

The rescoring module has been introduced in order to recover some recognition errors. The idea comes from some previous studies about syllabic speech recognition [5]. Basing on a corpus analysis in terms of the duration and energy of the syllables, some templates have been traced. Each unit template records the minimum and maximum value for energy and duration in the corpus. If a recognized syllable does not respect this ranges, then a set of static rules is checked. In this phase the system performs substitutions, insertions or deletions of units, on the basis of some systematic errors of the recognizer, or possible confusions between syllables, from the energy profile point of view.

Figure 2 depicts the Nuance based system. An environment setup is performed by an Abla module, which sets the noise type, the confidence level, the static grammar to be used (in Grammar Specification Language (GSL) format [6]), and prepares the signal for the recognition phase.

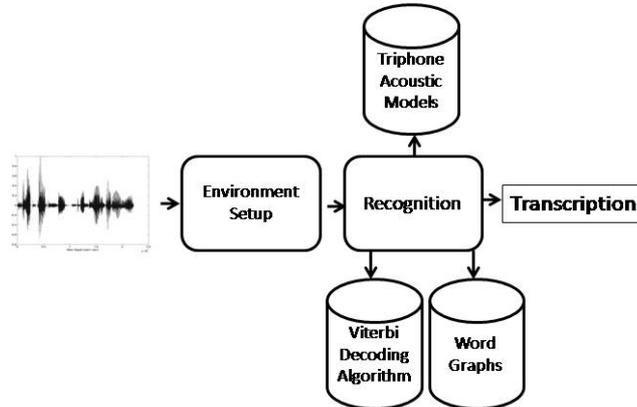


Fig. 2. Nuance ASR architecture.

The Nuance core uses standard Viterbi algorithm, and Word Graphs [4], in order to decode the speech signal, on the basis of triphonic acoustic models. A big quantity of training material was used to enroll the phonemes and then the triphones. Numbers cannot be reported because of company policy, but it's about two orders of magnitude higher than the TSpeech corpus.

### 3 Results

Table 2 shows the results of the two systems on the test data given by the EVALITA 2009 workshop for the *clean environment*.

Table 2. EVALITA 2009 performances of Abla srl vs. best ASR results in *clean environment*.

System Name	Sentence Accuracy	Unit Accuracy	Corr	Err	Del	Ins	Sub
Best ASR in EVALITA 2009	96.44%	99.45%	2350	13	8	3	2
Nuance Abla	95.89%	99.28%	2345	17	6	2	9
Abla TSpeech	81.64%	96.06%	2270	93	34	3	56

The TSpeech was trained on a small quantity of the training data from EVALITA 2009, because we didn't have time for segmenting all the material. As result, examples have been about 200 for each digit syllable.

Results show that the syllable is a promising unit to get acceptable performances in common applications, even if the training data is poor. Inside the process, the rescoring module rises the percentages as it recovers some systematic errors.

The Nuance solution gets very high performances, even if it is trained on a different training set respect to the one from EVALITA 2009. It is evident that good

models training and word graphs are able to take care of differences in speakers and recording environment.

It is to note that Nuance 8.5 is declared a state of the art system, even if its performances are not the highest in the EVALITA ranking. That is because, even if it is a highly trained system, it cannot be as good as a system trained on a corpus and tested on the same data. The training set from EVALITA poorly intersects the test data in terms of speakers, but recording environment and file formats are the same, and this makes the difference in our opinion, which explains the gap in performances respect to the best ASR.

**Table 3.** EVALITA 2009 performances of AblA srl vs. best ASR results in *noisy environment*.

System Name	Sentence Accuracy	Unit Accuracy	Corr	Err	Del	Ins	Sub
Best ASR in EVALITA 2009	87.77%	96.21%	3896	153	104	13	36
Nuance AblA	77.69%	88.65%	3604	458	268	26	164
AblA TSpeech	69.09%	82.23%	3375	717	467	56	194

Table 3 shows results on the noisy environment, respect to the best performing ASR. The most surprising result is that, even if the TSpeech was trained on clean environment recordings, it is able to recognize even in the noisy case. That is due to the rescoring module and to the usage of the syllable unit, which is more robust to changes in environment and speakers [2]. This is the real advantage of using syllable trained models.

Good performance is obtained by the Nuance system, even if only a clean environment setup was performed by AblA. The robustness of the algorithms and the variety of the training data on which the system was trained, are able to manage even high SNR ratio signals.

## 4 Discussion

We presented two approaches to the recognition of uttered digits sequences. The first one has a standard architecture, contained between two modules which are able to rise overall performances. Based on syllabic segmentation and acoustic models, even with poor training, the system gets acceptable performances, which can be useful in telephony applications. The most promising aspect lies in the usage of energy and duration templates inside a *rescoring* module, in combination to syllables, which are able to recognize speech even in noisy environment, and are less sensitive to pronunciation variations.

Surprising performances come from the Nuance system. Even if it is highly trained on company proprietary corpus, it is not able to get over those systems which have been trained directly on the EVALITA 2009 training corpus, distributed for this task.

A particular attention goes to the noisy environment performances of the two systems. They have not been trained on the noisy training data, but good performances are always found.

**Acknowledgments.** We would like to thank Santo Chianese, for his work in the implementation of the systems described in this paper, and Iolanda Alfano, who took care of the transcriptions at syllable level for ASR enrollment. Without their contribution this work would not have been possible.

## References

1. Abla srl, TSpeech speech recognizer for IVR solutions, [www.abla.it](http://www.abla.it)
2. Greenberg S.: Understanding speech understanding towards a unified theory of speech perception. Workshop on Auditory Basis of Speech Perception, ESCA (1996)
3. Nuance Automatic Speech Recognizer 8.5, speaker independent speech recognition engine for multichannel platforms, <http://support.nuance.com>
4. Huang, X., Acero, A., Hon, H.: Spoken Language Processing. Prentice Hall, New Jersey (2001)
5. Coro, G., Cutugno, F., Caropreso, F.: Speech recognition with factorial-HMM syllabic acoustic models, INTERSPEECH (2007)
6. Nuance GSL grammar specification, <http://cafe.bevocal.com/docs/grammar/gsl.html>