Connected Digits Recognition Using a Syllable-Based ASR System

Francesco Cutugno[1], Bogdan Ludusan[1], Antonio Origlia[1] and Serena Soldo[1,2*]

[1] NLP Group, Department of Physical Sciences, Federico II“ University,
Monte Sant'Angelo Campus, 80126 Naples, Italy
{cutugno, ludusan, soldo}@na.infn.it, antori@gmail.com
[2] IDIAP Martigny, Switzerland

Abstract. In this report we present the system proposed and the results obtained by our group for the Evalita 2009 Connected Digits Recognition task. The recognition system uses the syllable as base unit. In a first stage, the continuous speech sequence is divided in syllable-like units using an energy-based algorithm. Then, the obtained syllables are passed to a classifier in order to calculate the syllable/class probability distribution. In the final stage, a Viterbi-like decoding algorithm based on multistage graphs will find the most likely sequence corresponding to the audio input. The results obtained, 77.84% for digit accuracy on the test set, are not satisfying but our efforts are, at the moment, concentrated towards the optimization of this novel approach more than towards reaching state of the art results.

Keywords: Syllables, Support Vector Machines, Multistage Graphs

1 Introduction

John Garofolo, one of the most active members of NIST for speech and language evaluation campaigns, used to state [1] his preference for the systems that, participating at an evaluation, fell in the lowest part of the score list. While the top scorer often reaches state of the art results by inductively refining methods that have already shown their abilities and robustness, in the bottom part of the list sometimes (but not always...) original ideas are encountered and the real hope is that these ideas can lead in a not too distant future to systems able to give new directions to the research in the speech recognition field. We were aware that the system presented here would have provided very low performances, but we still decided to participate invoking the spirit and the inspiration coming from John's beliefs, beliefs that we share with him.

We have previously tested our system on a corpus of Italian numbers and we obtained, using the manually segmented syllables, 2.60% syllable error rate and 8.26% sentence error rate.

* The authors appear in strict alphabetic order

2 System Description

The architecture of our system is presented in Figure 1. A more detailed description of it is done in the following paragraphs.

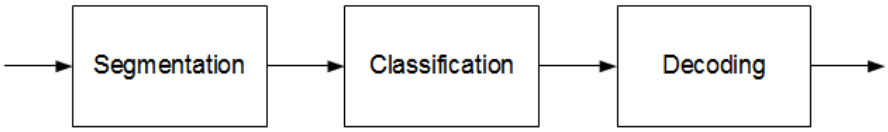

Fig. 1. Block scheme of our recognition system.

2.1 Segmentation

The syllable segmentation algorithm is based on the traditional energy based search [2], [3] for syllable boundaries which we improve by recognizing and treating separately a number of common situations. We also employ harmonicity analysis [4] and pitch based voice activity detection.

Before searching for syllable boundaries, the algorithm has to find the syllable nuclei. Since syllabic nuclei correspond to high energy regions, it is clear that our first goal is to look for energy peaks. Unfortunately, not every energy peak is a syllable nucleus: energy peaks can be caused, among other, by noise, tongue clicks, deep breaths or peculiar pronunciation and intonation strategies. We found that by adequately filtering and smoothing the energy profile it is possible to get rid of a number of artifact peaks, mostly caused by noise and in-between breaths. Although this strategy decreases the rate of false positives of the nuclei search process, it can not deal with artifact peaks that are pronunciation related and with peaks caused by particularly strong noisy events.

Before looking for syllable nuclei, another filtering step must be performed: in this step we search for recurrent patterns that cause a false positive to be detected by the syllable nuclei search step and filter them out. These patterns can be described as small peaks appearing on steep rising regions of the energy profile or as peaks falling in regions of the signal in which we observe abnormal pitch jumps.

With the majority of the artifact energy peaks filtered out, we can now perform syllable nuclei search on a relatively polished energy profile. Basically, for a peak to be accepted as a syllable nucleus candidate it must exhibit a certain degree of prominence. Another condition it has to satisfy is that of being voiced – in this sense we used a pitch-based voice activity detection to select only the voiced candidates. To improve the accuracy of our segmentation algorithm we chose to employ traditional voice activity detection and harmonicity analysis. Since syllable nuclei are harmonic, in general, we used the harmonic-noise ratio as an additional feature to consider when constructing the rule set to detect syllable nuclei.

The final step of the segmentation process is the syllable boundaries search. Syllable boundaries are generally located in the valleys preceding syllable nuclei, of the energy profile, but some exceptions arise when fricatives are involved. We have found that by analyzing the energy curve's slope before the valley and positioning the

syllable boundary marker where the slope reaches its maximum significantly improves the accuracy.

2.2 Classification

In order to compute the acoustic likelihood a model based on Support Vector Machines (SVMs) was used. An SVM is a supervised learning classifier and, in particular, it is a maximum-margin classifier. This means that during the training stage it tries to maximize the geometric margin of the separating hyperplane.

To construct the SVM classifier we used LIBSVM [5], a library for support vector classification able to perform multi-class classification. While LIBSVM provides several kernel functions, an experiment was conducted for deciding upon the kernel for the current task and the Radial Basis Function (RBF) kernel gave the best results. Our findings are in accordance with the claim [6] that the RBF kernel is the best suited for speech recognition.

We decided to represent the syllable as a static analysis unit in analogy to the process of feature extraction for face/image recognition algorithms. Thus, all internal variations due to coarticulation or, more generally, to acoustic-temporal dynamics, are merged into a set of features that the SVM classifier will consider as a whole. Our approach to representing the syllable uses a fixed number of frames per syllable, i.e. each syllable is described using a fixed number of frames, *n*, regardless of the syllable length. In order for these *n* frames to cover the entire syllable while keeping the length of the analysis frame constant, the shift between two consecutive frames had to be varied from syllable to syllable.

Figure 2 illustrates this approach, where the numbers from *1* to *n* represent the frame number, while the number under each frame represents the distance between the beginning of two consecutive frames (being the sum of the frame length and the shift between the current frame and the next frame). It can be seen that for shorter utterances the frames overlap, while for longer utterances there are gaps between two consecutive frames. For this task, the number *n* of frames per syllable employed in our representation was set to 21.

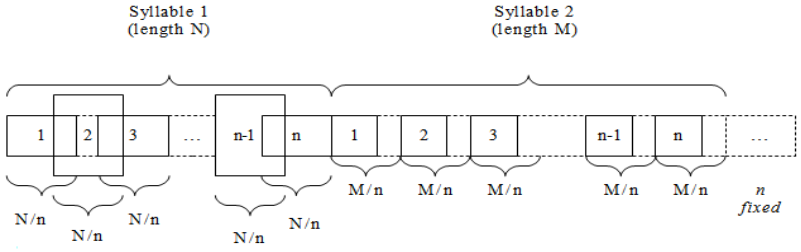

Fig. 2. The syllable representation employed in our system.

The feature used in the classification process are Mel Frequency Cepstral Coefficients (MFCC). In the feature extraction stage 13 MFCCs were extracted from every 16 ms analysis frames.

In the training stage, the SVMs used syllables obtained automatically from the Train set of the Connected Digits Recognition task. Due to the fact that the segmentation algorithm produces errors, we took into consideration only the utterances giving an equal number of syllables to the manual transcription.

2.3 Decoding

This classification system provides at its output information about the probability of each segment to belong to any of the possible classes. This information is used in combination with linguistic information to reconstruct the correct sequence of syllables. Technically, the information provided by the classification and those provided by the model language are described by a multistage graph (see Figure 3).

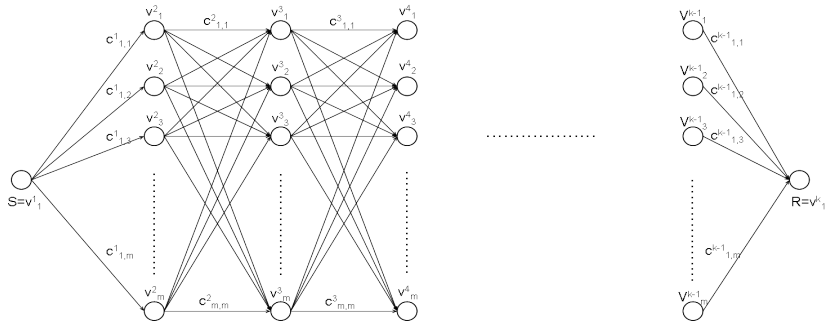

Fig. 3. Multistage graph structure for our decoding algorithm

The correct sequence of syllables is retrieved by a dynamic programming procedure computing the best path, based on the following recurrence equation:

$$C_i^l = \begin{cases} 0 & \text{if } l = i = 1 \\ max_j(\alpha \cdot C_j^{l-1} + \beta \cdot c_{j,i}^{l-1} + \gamma \cdot p_i^l) & \text{else} \\ \text{such that } j \text{ is the index of a node of level } l-1 & \end{cases}; \quad (1)$$

where:

C_j^{l-1} is the cost of the best path to reach the node v_j^{l-1};

$c_{i,j}^{l-1}$ is the cost of the edge between node j and node i (language model);

p_i^l is the score of the node i (acoustic model);

α, β and γ are the weights of the information coming from the acoustic model, the language model and the best path until that point respectively.

The cost of each edge of the graph is given by the bigram probability between the two corresponding syllables, computed on the training set. In addition to the bigram language model, a context-free grammar, having as terminal symbols the syllables, was employed. The grammar parser is used to select at each step of the recursion only the nodes that allow admissible paths.

This kind of structure could work very well in case of absence of segmentation errors. This perspective, unfortunately, is not realistic: any automatic segmentation

system produces some insertion or deletion errors that can make the decoding algorithm fail. For this reason the information present in the graph is enriched by adding new sets of edge (Figures 4 and 5) in order make the system robust in the case of segmentation errors. The new recurrence equation to define the best path on the graph is the following:

$$C_i^l = \begin{cases} 0 & if\ l = i = 1 \\ max \begin{pmatrix} max_j(\alpha \cdot C_j^{l-1} + \beta \cdot c_{j,i}^{l-1} + \gamma \cdot p_i^l), \\ W_{ins} \cdot max_j(\alpha \cdot C_j^{l-1} + \beta \cdot max_k(t_{i,j}^k) + \gamma \cdot p_i^l) \end{pmatrix} & if\ l = 2 \\ max \begin{pmatrix} max_j(\alpha \cdot C_j^{l-1} + \beta \cdot c_{j,i}^{l-1} + \gamma \cdot p_i^l), \\ W_{ins} \cdot max_j(\alpha \cdot C_j^{l-1} + \beta \cdot max_k(t_{i,j}^k) + \gamma \cdot p_i^l) \\ W_{del} \cdot max_j(\alpha \cdot C_j^{l-1} + \beta \cdot c_{j,i}^{l-2} + \gamma \cdot p_i^l) \end{pmatrix} & else \end{cases}; \quad (2)$$

where W_{ins} and W_{del} are the penalties for path with insertions and path with deletions respectively. These penalties are tuned based on the insertion and deletion error rates coming from the segmentation system.

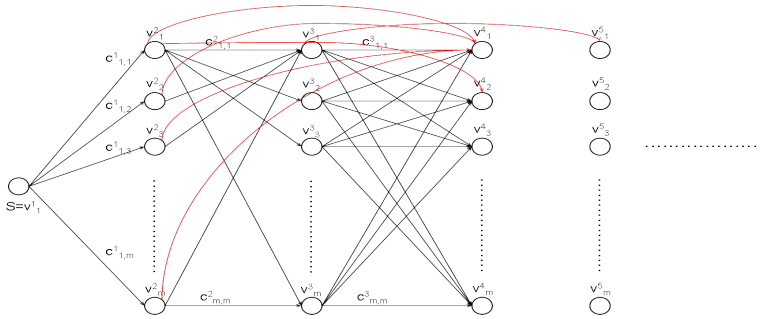

Fig. 4. Graph with edges to consider insertion errors

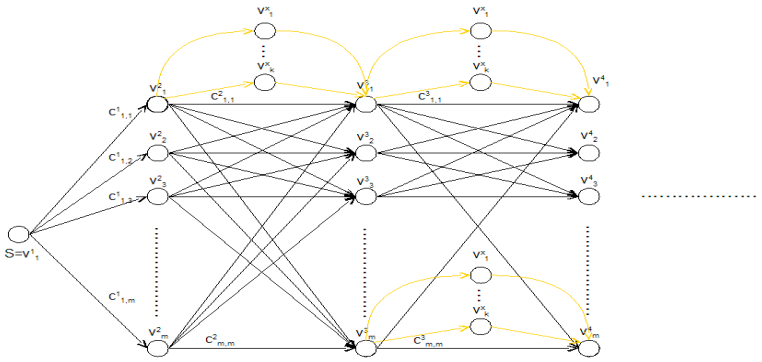

Fig. 5. Graph with edges to consider deletion errors

3 Results

The system was tested on the Test set of the Connected Digits Recognition task, consisting of 2360 digits from 365 utterances and the results obtained are presented in Table 1.

Table 1. Results obtained by the system.

	Total	Correct	Accuracy [%]
Words	2360	1941	77.84
Sentences	365	67	18.36

4 Discussion

At the time of writing this report it has not been possible yet to perform a detailed analysis of the results in order to understand what module of the architecture presented in Figure 1 had a bigger role in the errors produced by the system. Also, we cannot, at the moment, estimate the syllable segmentation procedure and classifier accuracy, but in general, the segmentation algorithm has an accuracy ranging from 75% to 82% and normally the decoding process confirms or even increases this performance. However, as it was shown in section 2.2 the classifier for this task has been trained on uncertain data, and this can explain, in principle, the low performance of the system. We will attempt to improve the results in the near future.

Acknowledgments. Bogdan Ludusan's work was supported by the EU FP6 Marie Curie Research Training Network “Sound to Sense”. Serena Soldo's work was supported by the European Union under the Marie-Curie Training Project SCALE.

References

1. Garofolo, J.: Evaluation of Language Technologies and Resources. Round table intervention, 3rd AISV National Conference. Trento (2006)
2. Petrillo, M., Cutugno, F.: A syllable segmentation algorithm for English and Italian. In: Proceedings of EUROSPEECH’03, pp. 2913--2916 (2003)
3. De Jong, N., Wempe, T.: Praat script to detect syllable nuclei and measure speech rate automatically. Behavior Research Methods vol. 41, issue 2, pp. 385--390 (2009)
4. Boersma, P.: Accurate short term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In: Proceedings of the Institute of Phonetic Sciences 17, pp. 97--110 (1993)
5. Chang, C. C., Lin, C. J.: LIBSVM: a library for support vector machines (2001)
6. Ganapathiraju, A.: Support Vector Machines for Speech Recognition. Doctoral Thesis, Department of Electrical Engineering, Mississippi State University (2002)