SYLLABLE ONSET DETECTION APPLIED TO THE PORTUGUESE LANGUAGE

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

Hugo Meinedo, João P. Neto and Luís B. Almeida

INESC - IST
R. Alves Redol, 9 1000-029 Lisboa - Portugal
Hugo.Meinedo@inesc.pt, jpn@inesc.pt, Luis.Almeida@inesc.pt
http://hebb.inesc.pt/NN/RFC

ABSTRACT

Recent developments have suggested that the use of syllables as the basic unit in a speech recognition system could be very usefull. Since syllable boundaries are more precise and well defined than phoneme ones there is a large scope for their application on the continuous speech recognition process. In this work we developed different methods of syllable segmentation in continuous speech. These methods are based on perceptually oriented feature extraction techniques. These features were post-processed through simple threshold mechanisms or by an artificial neural network based classifier in order to estimate the syllable boundaries. These systems were trained and evaluated using a Portuguese database with continuous speech. The results show that large context input windows ($260ms$) are the most appropriate, achieving results of 93% detection of onsets with insertion rates of only 15%.

Keywords: *Syllables, onset detection, artificial neural networks, continuous speech*

1. INTRODUCTION

Despite some great advances in automatic speech recognition (ASR) systems there still is a large scope for improvements. Under real world conditions such as with spontaneous speech, background noise and reverberation these systems still perform poorly. It is well know that speech is non-stationary and current frame-based ASR systems, where speech is modeled as a stream of conditionally independent frames, lack the ability to model speech dynamics realistically.

Recently we saw the introduction of systems that began to use syllables instead of phonemes as the basic unit for speech recognition. Some of these systems use syllables instead of triphones, reducing effectively the total number of units to model without an increase in word error rate [1].

It has been shown that explicit knowledge of syllable onsets can be useful on improving speech recognition systems [2]. Syllable boundaries are more precise and well defined than phoneme ones, so their detection has the potential to improve recognition by helping to accurately segment speech signals. This is a topic we are beginning to explore within the BD-PUBLICO database [3], our large vocabulary, continuous speech *corpus*.

Besides that, accurate estimation of syllabic boundaries can be incorporated into the decoding process limiting the number of hypotheses by restricting phonetic information to fit into syllable segments. Syllables as a whole appear to be more stable than the constituent phonemes so their explicit use may be helpful to minimize coarticulation problems.

Syllables have been used successfully in speech recognition systems for several languages, like the Chinese, Japanese and some European languages (Spanish and Hungarian). Being the Portuguese a highly syllabic language, we expect that the use of this kind of information will also introduce potential benefits in speech recognition tasks.

This paper describes our work in developing and evaluating methods which accomplish accurate syllabic segmentation of continuous speech, with the goal of in the near future integrating that information in our large vocabulary continuous speech recognition system for the Portuguese language.

On section 2 we present the different methods developed. The review of the speech database used during the training and evaluation procedures is presented in section 3. On section 4 training details are explained. On section 5 we summarize the results for each method. Finally, we will present our conclusions and future work on sections 6 and 7.

2. DEVELOPMENT OF DIFFERENT METHODS FOR SYLLABLE ONSET DETECTION

The following methods attempt to accurately detect the syllable boundaries, given an input continuous speech signal. They accomplish this task by some form of perceptually oriented feature extraction and their consequent post-processing.

2.1. First method

This method is based on the work of Wu *et al* [4] where they present a set of features based on the analysis of energy trajectories in critical band channels. Large values in this representation correspond to patterns of synchronized rises and falls in subband

energy regions where syllable onset characteristics occur. These syllable-length intervals are on the order of 100-250ms, matching closely the duration of most syllables. We made a slight modification in the half-wave rectification stage to preserve also the negative changes in energy. These constitute an additional set of 9 critical band coefficients. These features were used to directly assert syllabic onset locations. For each time frame, the coefficients from the 9 critical band channels that preserve positive changes are summed up and the result is normalized. The same is done for the 9 negative changes coefficients. A syllable mark is then estimated applying a simple mechanism of thresholds to this time set of features (positive and negative).

2.2. Second Method

In this second method we replaced the heuristic threshold mechanism of previous method by a neural network classifier, where the critical band channel features constitute the input frame. The classifier consists of a fully connected, feed-forward, multilayer perceptron (MLP) with a single-hidden layer and a single output unit. This MLP was trained through the backpropagation algorithm to estimate the probability that a given frame of input features is a syllable boundary. Additionally we used several adjacent input frames to add some acoustic context.

2.3. Third Method

The next approach was to additionally include perceptual linear prediction (PLP) coefficients at the input of the classifier. This analysis method is the one used in our large vocabulary speech recognition system. With this addition each input frame results from the concatenation of these PLP coefficients with the same features that we used for the previous method.

2.4. Fourth Method

Finally, and for comparison purposes, we tested the system with just the PLP parameters as input to the neural network classifier. This configuration of feature extraction is similar to the one employed by our large vocabulary speech recognizer. This way we were able to analyze the importance of mixing together different sources of information, and to discern their relative importance when used in a particular task.

3. DATABASE

These methods were trained using a subset of the EUROM.1 SAM Portuguese *corpus* [5]. This database consists of read continuous speech, recorded in a sound proof room. It contains three different sets of speakers and different recording material. We selected for training and evaluation the *passages* of the so called Few Talker Set (10 speakers with 15 passages each, giving a total of 150 passages). In this subset, there were 40 different passages. Each passage is composed of 5 thematically connected sentences, giving a total of 750 sentences. Each speech file contained the whole passage. The database contains a total of 3,408 words, of which 1,314 were different words and 616 different syllables. It is a small *corpus* but has the advantage of being manually phonetic segmented.

4. TRAINING

The lexicon with phone transcription was converted automatically to syllabic information, using programs specially developed. This information was then aligned with the phonetic segments of the training speech files creating syllable segmentations. Furthermore, the *corpus* subset was phonetically hand labeled by linguists, avoiding the need for a phonetic forced alignment procedure when we automatically determined syllable boundaries.

The different feature extractions performed by the methods are all updated every 10ms, being the PLP features computed over a 20ms window and the analysis of energy trajectories in critical band channels over a 64ms window.

4.1. Threshold mechanisms

In the first method the two vectors of coefficients, produced by averaging the positive and negative 9 critical band coefficients, are used in conjunction with two numerical thresholds to produce an estimate of the syllable onset location by considering a syllabic mark the average point of the two marks produced by each threshold.

In the methods that use the neural network classifier, the syllable boundary estimate is produced by numerically thresholding the output probability generated by the classifier. Several tests were conducted to assert the best threshold values.

4.2. PLP analysis

A PLP feature extraction is applied to the speech signal. From this pre-processing phase results a frame with the log energy and PLP-12 cepstral coefficients and their first temporal derivatives. Therefore the feature vector has a total of 26 coefficients.

4.3. MLP training

The core of the last three methods, is a multilayer perceptron trained with the backpropagation algorithm. To improve the generalization capabilities of the algorithm we use a cross-validation set made with 150 of the 750 sentences present in the *corpus* subset. The remaining 600 sentences were chosen for training.

The desired output vector for the MLP training is composed of zeroes for the frames not corresponding to syllable boundaries and two consecutive ones indicating a syllable mark. This mark spans through 30ms, due to the 10ms superposition of frames. Some experiments were initially conducted to determine the correct position for these two frame marks. These experiments revealed that the best configuration corresponded to the frame of the actual onset and the

next contiguous one. This training situation is very demanding for the MLP because the output unit is only active a very small fraction of the time and it tends not to train the syllable boundaries. The solution was to weight the back propagation error produced at the output according to the desired label [7].

5. EVALUATION OF THE DIFFERENT METHODS

Although the MLP has been trained with just two frames marking syllable boundaries, the results were evaluated using a more flexible method. After obtaining the syllable boundary estimates, an evaluation algorithm checks if a mark falls within a window composed of 7 contiguous frames, which serve as a tolerance time interval of $80ms$, defined around the original syllable mark. If the syllable boundary estimate is within this window it is considered as correctly detected. Otherwise we consider that we were not able to detect that boundary. This has a potential disadvantage, because if a boundary occurs outside this tolerance window the algorithm will score a deletion (no mark inside the window) and score an insertion. The results present in all the tables are expressed as percentages of correctly marked boundaries and percentage of inserted boundaries.

5.1. First Method

With this method we tested a set of features based on the analysis of energy trajectories in critical band channels and a mechanism of thresholding to estimate syllable boundaries. The results obtained are presented in Table 1. The results were not satisfactory due, in a first analysis, to the simple threshold mechanism that was applied to decide the syllable onsets locations.

% correct boundaries	% inserted boundaries
55.18	35.58

Table 1. Results for the first method.

5.2. Second Method

Through this method we used a MLP classifier to estimate the syllable boundaries instead of the simple threshold mechanism of the previous method. Initially the number of hidden units of the MLP was fixed at 500 units, and we conducted tests to determine the number of contiguous context frames at the MLP input that produced the best results. These experiments are summarized in Table 2.

The correct boundaries detection improved when we increased the input context window despite the increase in the number of insertions. The last entrance in Table 2 shows what happens when the number of hidden units increases. In this case there is no direct benefit with more hidden units on the MLP.

MLP configuration	% correct boundaries	% inserted boundaries
(13x18)-500-1	54.94	12.10
(15x18)-500-1	55.78	11.35
(19x18)-500-1	57.43	11.57
(23x18)-500-1	59.02	12.08
(25x18)-500-1	63.81	14.44
(25x18)-1000-1	76.68	30.42

Table 2. Results for the second method. MLP configuration is: (frames in context window x features in one frame of input) - hidden units - output units.

Although results have improved when compared to the first method, they remain far from satisfactory.

5.3. Third Method

In previous methods we were using just the set of features based on the analysis of energy trajectories in critical band channels. In this method we additionally use the PLP coefficients maintaining the MLP classifier. Again a series of tests were conducted to determine which MLP configuration produced the best results. In Table 3 we evaluated the performance of the method for different number of hidden units. We observed a significant improvement compared with previous methods. However there was also an increase of the insertions.

MLP configuration	% correct boundaries	% inserted boundaries
(19x44)-200-1	84.72	23.59
(19x44)-400-1	83.70	23.91
(19x44)-500-1	84.16	22.41
(19x44)-1000-1	84.27	22.02

Table 3. Results for the third method, maintaining the number of input parameters constant.

From these results we chose the configuration with 200 units in the hidden layer and increased the number of frames of context to 25. This (25x44)-200-1 configuration achieved 94.41% of correctly detected syllable boundaries and mistakenly inserting 18.72%.

5.4. Fourth Method

This last method was only based on the PLP parameters discarding the energy trajectories in critical band channels features. We started with the same MLP hidden layer size as in the previous method, 200 hidden units, varying the number of input frames. The results are presented in Table 4.

With 200 units we do not reach the performance obtained by the previous method. But here we are using less parameters in the MLP. When we raised the hidden units to 300, similar performance was obtained (Table 5). Also, we observed that there was no gain increasing the context window beyond 25 frames.

MLP configuration	% correct boundaries	% inserted boundaries
(17x26)-200-1	91.48	20.80
(21x26)-200-1	92.34	20.18
(25x26)-200-1	92.96	19.79

Table 4. Results for the fourth method, with 200 hidden units.

MLP configuration	% correct boundaries	% inserted boundaries
(21x26)-300-1	93.22	17.59
(25x26)-300-1	93.53	14.70
(29x26)-300-1	92.98	15.23

Table 5. Results for the fourth method, with 300 hidden units.

If we compare these results with the ones from previous methods, specially the third, we see that with less parameters, a somewhat similar result was obtained, 93.53% correct boundaries detected and only 14.70% insertions compared to the 94.41% and 18.72% for the third method.

5.5. Variation of the tolerance window

All the previous results were obtained with a 7 frame tolerance window. We conducted an experiment with more stringent tolerance windows, composed of 5 frames ($60ms$) and 3 frames ($40ms$). We use the previous method with the MLP configuration of (25x26)-300-1. In Table 6 we see that the performance is not substantially affected for the 5 frames ($60ms$) case, but degrades more for the 3 frames ($40ms$) case, primarily due to the deletion+insertion side effect.

tolerance frames	% correct boundaries	% inserted boundaries
7 ($80ms$)	93.53	14.70
5 ($60ms$)	92.04	16.11
3 ($40ms$)	86.94	21.21

Table 6. Results with different tolerance windows.

6. CONCLUSIONS

The results show the importance of the MLP for a robust estimation of the syllable boundaries by correctly detecting more than 93% of the onsets and mistakenly inserting less than 15% where there were none. We saw that larger context windows are required for capturing syllabic information. The best results were achieved with 25 frames ($260ms$) of context while phoneme systems seldom pass the 9 frames barrier. We also conclude that although PLP parameters were developed for smaller units like phonemes, it does not invalidate their successful use in tasks with larger units.

7. FUTURE WORK

The next step will be to incorporate this information into our LVCSR system, that was trained using the BD-PUBLICO database [3]. There are several potential uses for it:

- Helping in the forced alignment process, by providing alignment marks.
- Reducing the word error rate during the recognition, by incorporating synchronization syllable boundary information into the decoding process.

8. ACKNOWLEDGMENTS

This work was partially funded by the PRAXIS XXI project 1654. An acknowledgment is also given to Prof. Isabel Trancoso to make available the SAM database.

REFERENCES

[1] J. Hamaker, A. Ganapathiraju, J. Picone, J. Godfrey, *Advances in Alpha Digit Recognition Using Syllables*, in Proceedings ICASSP 98, Seattle, USA, 1998.

[2] G. Cook, T. Robinson, *Transcribing Broadcast News with the 1997 ABBOT System*, in Proceedings ICASSP 98, Seattle, USA, 1998.

[3] J. Neto, C. Martins, H. Meinedo and L. Almeida, *The design of a Large Vocabulary Speech Corpus for Portuguese*, in Proceedings EUROSPEECH 97, Rhodes, Greece, 1997.

[4] S. Wu, M. L. Shire, S. Greenberg, N. Morgan, *Integrating Syllable Boundary Information into Speech Recognition*, in Proceedings ICASSP 97, Munich, Germany, 1997.

[5] C. Ribeiro, I. Trancoso and M. Viana, *EUROM.1 Portuguese Database*, Report of ESPRIT Project 6819 SAM-A, 1993.

[6] J. Neto, C. Martins and L. Almeida, *The Development of a Speaker Independent Continuous Speech Recognizer for Portuguese*, in Proceedings EUROSPEECH 97, Rhodes, Greece, 1997.

[7] S. Lawrence, I. Burns, A. Back, A. Tsoi and C. Giles, *Neural Networks Classification and Prior Class Probabilities*, in Tricks of the trade, Lecture Notes in Computer Science State-of-the-Art Surveys, Springer Verlag, pp. 299-314, 1998.