THE USE OF SYLLABLE SEGMENTATION INFORMATION IN CONTINUOUS SPEECH RECOGNITION HYBRID SYSTEMS APPLIED TO THE PORTUGUESE LANGUAGE

Hugo Meinedo *João P. Neto*

INESC - IST
Rua Alves Redol, 9 1000-029 Lisboa - Portugal
Hugo.Meinedo@inesc.pt, jpn@inesc.pt
http://neural.inesc.pt/NN/RFC

ABSTRACT

Recent works have showed that the use of syllables as the basic unit in a speech recognition system could be very useful. These works introduced methods exploiting syllable information as a mean to add robustness in "traditional" systems that use phonemes/phones as the basic unit. Being the Portuguese a highly syllabic language we expected that information from syllables would introduce potential benefits in speech recognition tasks. Following these ideas we started by developing different methods of automatic syllable segmentation. Next we applied the best segmentation method to our large vocabulary continuous speech corpus (BD-PUBLICO) achieving an accuracy of 72%. We developed a process to use the segmentation information in the acoustic models of our baseline speech recognisers for the Portuguese language. The results obtained by the modified recognition systems on 5k and 27k vocabulary tasks showed that the use of basic syllable segmentation information helps the systems to improve their overall performance by roughly 10%.

1. INTRODUCTION

Over the last years researchers have been trying to improve automatic speech recognition performance under real world conditions by using more knowledge of how humans recognise speech. Researchers had found numerous leads suggesting that the syllable may be perceptually very important because the human speech perception mechanisms make extensive use of temporal information, broader than the phoneme which stands today as the most used basic unit for automatic speech recognition [1, 2]. Also, several prosodic features are easier modelled using syllables [3]. Besides that syllable boundaries are more precise and well defined than phoneme ones, assuming that we clearly define the syllable concept and segmentation rules, so their detection has the potential to improve recognition.

Following these ideas we started to work on syllable segmentation methods [4]. Our main goal was to use the syllable segmentation information in the acoustic model of our continuous speech recognition hybrid MLP/HMM system (in this system the acoustic model is implemented through a Multilayer Perceptron classifier). We started by using that information (the existence or not of a syllable boundary) at the input of our acoustic model as an additional feature to the coefficients and derivatives resulting from a standard PLP, Log-RASTA or ModSpec analysis. These feature extraction methods have significantly different temporal properties and the first two have been widely used in phone based systems. With the addition of syllable segmentation information we intend to add time synchronisation information alleviating the poor definition of phone boundaries, searching for better recognition performance and reduction of recognition time.

In this paper we describe the work done in the development of our syllable segmentation system applied to our large vocabulary continuous speech corpus, the BD-PUBLICO. After obtaining a good segmentation performance we developed and evaluated a method for incorporating the syllable information, coming from the segmentation system, in our baseline speaker independent continuous speech recognition systems that have been developed for the Portuguese language.

On section 2 we present the syllable segmentation method used through our work. The process for incorporating the syllabic information in the speech recognition system is described on section 3. On section 4 the evaluation of the modified recognition systems is summarised. Finally, some conclusions and future work are presented in sections 5 and 6.

2. SYLLABLE SEGMENTATION METHOD

In previous work [4] we developed different methods of automatic syllable segmentation applied to a small Portuguese continuous speech corpus of read speech but with the advantage of being hand phone labelled. Based on this hand phone segmentation and a set of syllable construction rules we were able to achieve an accuracy of 79% on the syllable segmentation task.

This syllable segmentation system, whose block diagram is represented in Figure 1, uses PLP-12 cepstral coefficients,

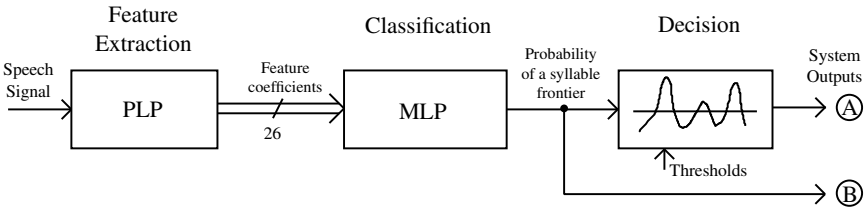

Figure 1: Syllable segmentation system.

log-energy and their first temporal derivatives as feature extraction every 10 ms applied to a 20 ms frame. These 26 feature coefficients are fed to a MLP classifier that proved essential for obtaining a robust estimation of the syllable boundaries. The MLP incorporates local acoustic context via a 25 frame input window (12 frames of left and right context around the central frame). The resulting network has a single hidden layer with 300 units and 2 complementary outputs representing the probability of a given frame having a syllable frontier or not. The last step receives the MLP output and applies a threshold detector and a simple decision rule. This algorithm searches the peaks in the probability output aided by two thresholds and marks the first frame where the peak occurred as a syllable boundary. The output of this thresholding stage (marked as **A** in Figure 1) assumes only two values, "1" if the input frame has a syllable boundary and "0" if not. Normally, this is the output used when the syllable segmentation method is evaluated. After some preliminary experiments we verified that the direct MLP output (marked as **B** in Figure 1) could also be very useful.

Next we tested this segmentation method, that had been trained on the previous database, on our large vocabulary continuous speech corpus, the BD-PUBLICO [5], with very poor results. The main reasons found for this were: i) The average maximum speech signals amplitude varies greatly between the two databases, which made the segmentation system very sensitive to small noises, causing many insertions; ii) In the BD-PUBLICO there are many more different syllables, 2,727 against 1,161 in the previous small training database, which inevitably produced higher confusability between similar classes and lead to classification errors; iii) The desired outputs in all of the BD-PUBLICO sets are based in an automatic forced alignment of ideal transcription of words while in the small database the subset used for training was hand labelled by linguistic personnel based in what really exists in the speech files. This was a limiting factor and will always induce a somewhat lower classification performance even when the systems were trained directly over the BD-PUBLICO database.

After this it was clear for us that we could not use directly in BD-PUBLICO the syllable segmentation system trained on the previous small database. The next step was to develop a new syllable segmentation system trained directly on one of the training sets of the BD-PUBLICO, based on labels resulting from an automatic phonetic alignment process and using the same set of syllable construction rules previously developed. Another modification was to double the number of hidden units in the MLP to cope with the increase in the number of different syllables. This larger MLP caused a very slow training (about one full day for each epoch). The solution found was to create a smaller training set having only 1/8 of the total number of frames present in the complete set. The MLP was first trained in this smaller set and after it had achieved a good frame level classification rate, the training changed to the larger set. Table 1 summarises the results obtained on these two situations evaluated on the same set of BD-PUBLICO. The first one represents the MLP trained exclusively with the smaller set (1/8) and the second one after re-training in the larger set (Complete). In this case we started with the MLP weights that had obtained the best frame level classification in the 1/8 set. It is indicated the percentage of correctly detected syllable boundaries, the percentage of inserted boundaries where there were none and the percentage of Accuracy which represents the correctly minus inserted boundaries.

Training sets	Evaluation in BD-PUBLICO		
	% Correct	% Insertions	% Accuracy
1/8	79.7	14.0	65.7
Complete	84.6	12.8	71.8

Table 1: Evaluation results for the syllable segmentation system with an MLP having 600 units trained first on the 1/8 set and afterwards on the Complete training set of the BD-PUBLICO.

By re-training in the total number of training patterns we were able to achieve better segmentation results not only increasing the percentage of correctly detected syllable boundaries but also decreasing the number of insertions. This is a very good result when compared to our initial system trained and tested in the previous database, where we got an Accuracy of 79% since we are now training and evaluating the syllable segmentation system based on automatic phonetic alignment generated labels.

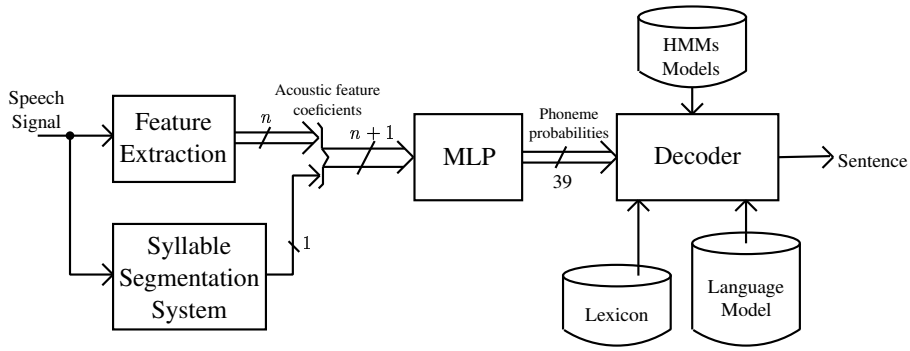

Figure 2: Modified speech recognition system using the information from the syllable segmentation system.

3. USE OF SYLLABLE INFORMATION IN SPEECH RECOGNITION

After the development of a syllable segmentation system for the BD-PUBLICO database we integrated that information in our hybrid MLP/HMM continuous speech recognition systems for the Portuguese language (See Figure 2).

We started with one of ours baseline recognition systems using PLP feature extraction and the MLP with an architecture of (7x26)-1000-39, that is, 7 input frames each one with 26 parameters derived from the PLP analysis, 1000 hidden units and 39 output units, corresponding to the phonemes used for the Portuguese. The information coming from the syllable segmentation system is used as an additional coefficient at each frame of input parameters to this MLP (See Figure 2). It was necessary to train this new MLP due to the additional parameter.

Initially we performed two experiments in order to evaluate what type of information produced by the segmentation system was more useful to the recogniser. In experiment **A** the additional coefficient coming from the syllable segmentation system is taken at the output A, see Figure 1, that comes from the decision stage and assumes only the values "1" if the input frame has a syllable frontier and "0" if not. In experiment **B** the additional coefficient is taken at the output B, more precisely the output of the MLP stage that varies between 0 and 1 and indicates the probability of a given frame being a syllable boundary.

5k vocabulary task	% WER
Experiment using output **A**	14.7
Experiment using output **B**	13.5

Table 2: Results for the two experiments using alternatively each output of the syllable segmentation system.

In Table 2 we compare the evaluation of both systems on a 5k vocabulary task. The syllable segmentation system used was not the final version but the one trained in the smaller (1/8) sized training set. We see that experiment **B** achieves a lower word error rate compared with experiment **A** since the recogniser's MLP learns to extract useful information from the signal that represents the probability of having a boundary in the input frame, provided by the segmentation system, while in experiment **A** that information may be lost due to the heuristic threshold procedure.

4. RESULTS IN SPEECH RECOGNITION

After these preliminary experiments the best syllable segmentation system was applied to three of ours baseline speech recognisers having acoustic models with different methods of feature extraction and all having the MLP using 7 input frames, 500 hidden units and 39 output units. They were all trained using the BD-PUBLICO database and evaluated using a development test set applying a 5k or a 27k vocabulary. The evaluations were based on two parameters: percentage of word error rate and decoding time expressed as the number of times the real time (xRT).

The first baseline recognition system uses log-energy and PLP-12 cepstral coefficients and their first temporal derivatives summing up to 26 acoustic parameters per frame. The second baseline system uses Log-RASTA analysis instead of the PLP having the same set of 26 coefficients. Finally, the last recognition system was developed with Modulation Spectrogram (ModSpec) features [6]. This feature extraction process outputs 28 coefficients in each frame resulting from the ModSpec analysis.

Table 3 summarises the results obtained in the evaluation on a 5k and a 27k vocabulary task for the baseline and the modified PLP recognition system.

This technique of using information from a syllable segmentation system within the acoustic model of a speech recogniser resulted in a decrease of word error rate superior to 6% relative for practically the same decoding time. This was a very encouraging result and proved us the utility of syllable information in continuous speech recognition tasks. The evaluation results achieved for the second baseline and the modified recognition system using log-RASTA analysis are summarised in Table 4.

System	5k vocabulary		27k vocabulary	
	% WER	x RT	% WER	x RT
Baseline	14.2	2.3	15.8	4.4
Baseline+Sil	13.3	2.1	14.7	3.6

Table 3: Results for the baseline recognition system with PLP features and for the modified system using syllable segmentation information.

System	5k vocabulary		27k vocabulary	
	% WER	x RT	% WER	x RT
Baseline	13.7	4.1	15.7	10.6
Baseline+Sil	12.3	2.7	14.2	6.1

Table 4: Results for the baseline recognition system with Log-RASTA features and for the modified system using syllable segmentation information.

Once again the use of syllable information provided a better recognition result with a 10% relative decrease in word error rate and a significant 37% relative decrease in decoding time. Finally, the results for the evaluation of the baseline and modified recognition system using ModSpec analysis are presented in Table 5.

System	5k vocabulary		27k vocabulary	
	% WER	x RT	% WER	x RT
Baseline	16.2	5.2	17.7	12.1
Baseline+Sil	16.3	3.5	18.8	8.3

Table 5: Results for the baseline recognition system with ModSpec features and for the modified system using syllable segmentation information.

In this case, although we observed an increase in the classification rate at the frame level, both in the training set and in the cross validation set, the modified system only improved substantially the decoding time but had an increase in the word error rate. A plausible explanation for this is that the ModSpec features are known for enhancing syllable nuclei [6] while the information provided by the additional coefficient from the syllable segmentation system responds to syllable frontiers. Both of them are providing information about syllable time scales but this information might have originated a mismatch that in the end did not translate into a better recognition.

5. CONCLUSIONS

In this paper we started by presenting the development of a syllable segmentation system for a Portuguese continuous speech *corpus*. This system was based in our previous work where we evaluated different syllable segmentation methods [4]. Next, we developed a method for integrating the syllabic information derived from the segmentation system in our baseline speech recognisers. Different experiments were performed to determine the best syllabic information in order to boost the gain in the speech recogniser. This technique was applied to our three baseline continuous speech recognition systems based on different feature extraction processes. The results obtained show improvements in word error rate and decoding time especially with the recognisers that use phoneme oriented features, like the PLP or Log-RASTA.

6. FUTURE WORK

These results open up new perspectives for the use of syllabic information within speech recognition. We are still devising new approaches to the incorporation of syllable segmentation information on speech recognition systems in order to boost performance. Besides that other potential uses for syllable segmentation information are being investigated namely in helping in the forced alignment process, by providing temporal marks which will reduce the uncertainty specially in the first iterations of the alignment process.

7. ACKNOWLEDGEMENTS

This work was partially funded by the PRAXIS XXI project 1654. An acknowledgement is also given to the PÚBLICO newspaper for making its texts available for this work.

8. REFERENCES

1. Greenberg, S., *On the origins of speech intelligibility*, in Proceedings ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels, Pont-a-Mousson, France, 1997.

2. Massaro, D., *Perceptual units in speech recognition*, Journal of experimental Psychology, 102 (2):199–208, 1974.

3. Lea, W., Medress, M., and Skinner, T., *A prosodically guided speech understanding strategy*, IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-23(1):30–38, 1975.

4. Meinedo, H., Neto, J., and Almeida, L., *Syllable onset detection applied to the Portuguese language*, in Proceedings EUROSPEECH 99, Budapest, Hungary, 1999.

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

6. Kingsbury, B. E., Morgan, N., and Greenberg, S., *Robust speech recognition using the modulation spectrogram*, Speech Comunication, 25:117–132, 1998.