Wavelet Transforms for Non-Uniform Speech Recognition Systems

Léonard Janer, Josep Martí, Climent Nadeu
Dept. TSC Universitat Politècnica de Catalunya
08034 Barcelona, Spain
Email: leonard@gps.tsc.upc.es

Eduardo Lleida-Solano
GTC Dept. IEEC Centro Politécnico Superior de Ingenieros
50015 Zaragoza, Spain
Email: lleida@mcps.unizar.es

ABSTRACT

A new algorithm for *non-uniform speech segmentation* and its application in *speech recognition* systems is presented. A new method based on the Modulated Gaussian Wavelet Transform based Speech Analyser (MGWTSA) and the subsequent Parametrization block is used to transform a uniformly signal into a set of non-uniformly separated frames, with the accurate information to be fed to our *speech recognition* system. Our algorithm wants to have a frame characterizing the signal where it is necessary, trying to reduce as much as possible the number of frames per signal, without an appreciable reduction in the recognition rate of the system.

1. Introduction

In the last years, Wavelet Transform (WT) have been applied in different speech processing applications[4] [3] as an efficient front-end system taking advantages of their good time-frequency resolution. Most of those systems are speech coding systems [2][7] or pitch detection systems [6][1][8]. Even though some speech recognition systems based on WT have been designed and tested [5], none of them tries to work with non-uniform parameters, as we are doing.

The work [1] we present in this paper involves speech parametrization using WT and speech recognition systems using Hidden Markov Models (HMM). The first step is the parametrization: the speech signal is analysed using a Modulated Gaussian Wavelet transform analyser with 17 bands (scales) distributed on a Bark scale [6]. In this first step the signal is decomposed into 17 different temporal signals, each one with a different frequencial information, as they are decompositions of the input speech at 17 scales. Actually these generated signals are taken sample by sample, but in a near future the system will work with a less accurate precision.

Once the signal is treated, we will examine the output of the analyser to detect instants of relevant information in the input, then we will take a frame at this time (composed by the 17 scales output samples) and finally, we will send it to the recognition system.

In the second section we will explain the segmentation algorithms, that resolve which are those relevant instants of information. In the third section, the recognition models will be presented and in the following one results of some tests will be shown to appreciate the performance of both the segmentation algorithm and its application in a speech recognition task. In the last section the conclusions of this work will be detailed.

2. Speech Segmentation using Wavelets

The segmentation step of our algorithm tries to detect relevant points in the signal. The solutions we present in this paper work basically with the information on two of the 17 scales of the ouput of the Modulated Gaussian Wavelet Speech Analyser. The two ones selected are number 3 and 9 (central frequency around 350-450 Hz. and 1170-1370Hz. respectively). In the following paragraphs the different solutions are detailed, and their segmentation process shown in Figures 1 and 5 for model 1, Figures 2 and 6 for model 2 and Figures 3 and 7 for model 3, both for non-connnected and connected digits.

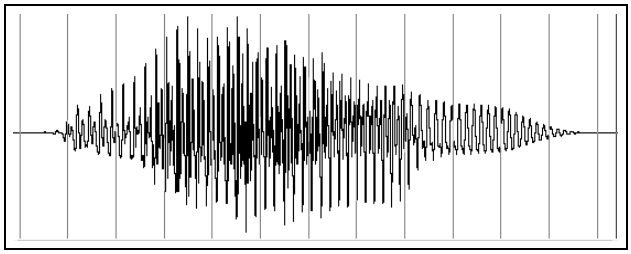

Figure 1: Uniform Segmentation with interframe distance equal to 300 samples for the digit "1" male speaker "ae"

1. **Uniformly separated frames:** We take a frame with a constant time interval along the signal. This will be our reference model, for which the evaluation is shown in Figure 4 and in Table 1, in the case of Non-Connected Digits.

[1] This work has been supported by the Spanish Ministery of Education and Sciences (MEC) grant TIC95-0884-C04

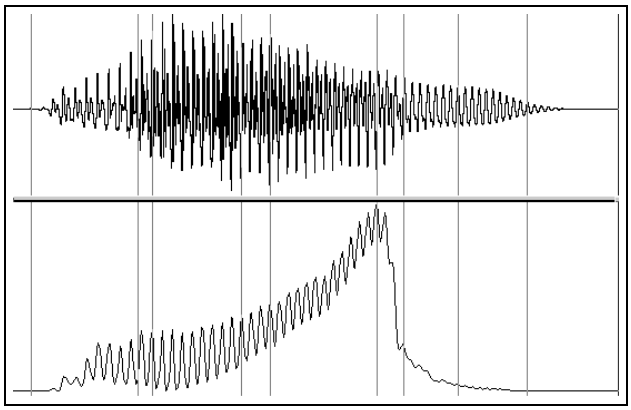

Figure 2: Non-Uniform Segmentation based on maxima location for the digit "1" male speaker "ae". Speech Signal, and output from the third band of the MGWTSA

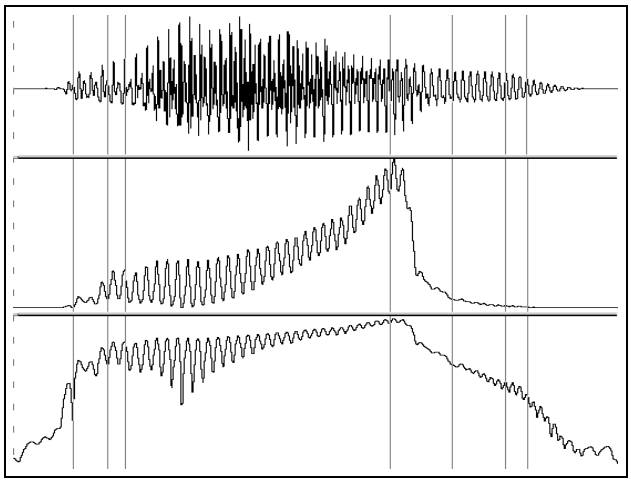

Figure 3: Logarithmic Non-Uniform Segmentation for the digit "1" male speaker "ae". Speech Signal, and output from the third band of the MGWTSA, and logarithmic detection signal

2. **Scales 3 and 9 Maxima Detection:** We select the frames looking only to those two scales. Basically we consider the samples beyond a threshold, and then iteratively select the maxima of the two bands, in a predetermined time interval. In Tables 3 and 4 the recognition evaluation for this model is displayed for different maxima postprocessing systems.

   In Segmentation Model *"M1"* for the two bands we compute the threshold, and select as much as one maximum per each 60 samples. Then, the two bands maxima positions are joined together ensuring that there is no two maxima at less than 55 samples. In Segmentation Model *"M2"* the system is the same, but the distance between two consecutive maxima is increased in order to obtain a fewer number of frames in average per digit. The Model *"M3"* takes each 5000 samples no more than 10 maxima (the peaks are iteratively eliminated to get this condition), as a way to ensure a minimum averaged interframe distance of 500 samples per scale, that is slowly reduced when the two bands are joined. For the Models *"M4"*, *"M5"* and *"M6"* we eliminate progressively one maximum each two or three maxima in an increasing time interval.

3. **Logarithmic solution:** It is the same idea as in the second solution, but we take the logarithm of the two scales as a previous step of the algorithm, to smooth the dynamic range of the signal, and then avoid some problems with the initial threshold (see the marks set in Figures 6 and 7). Results on our Speech Recognition Experiments are shown in Table 2.

Recognition Rate for Uniform Segmentation and increasing Interframe Distances	
Interframe Distance	*Recognition Rate*
80	86.7
120	85.15
200	83.2
250	82.5
300	82.5
350	78.2

Table 1: Speech Recognition Rate for Uniform Segmentation

Recognition Rate			
Recognition Rate	*Averaged Interframe Distance*	*Averaged No. of Frames per Digit*	Model Number
88.23	434	8.8	*"L1"*
77.9	1316	2.9	*"L2"*
81.17	603	6.3	*"L3"*
87.99	491	7.8	*"L4"*
87.82	516	7.4	*"L5"*

Table 2: Speech Recognition Rate for Logarithmic Non-Uniform Segmentation and two Gaussian Mixtures per Model

In Model *"L1"* the process is the same as in Model *"M1"* but working with the logarithm of the maxima. In the Model *"L2"* once the maxima are found we integrate the values around them and then localize the maxima of the transitions of the integral, up to the mininum number of frames possible (the averaged number of frames per digit is below 3). In Models *"L3"* and *"L4"* we apply the same process than in the latest case but the number of maxima is not so drastically reduce. Finally, in Model *"L5"* the output of bands 3 and 9 from the MGWTSA is only computed each 10 samples, as a way to reduce the computational load of the system, without an appreciable recognition rate reduction (the convolution is performed once each 10 samples).

In all the alternatives of our algorithm the objective is always the same: avoid as much as possible irrelevant information to the speech recognition step, as silent speech segments would be. In the results, we always indicate the averaged number of frames per digit and the averaged number of samples between two consecutive frames, as an idea to compare our results with uniformly segmented algorithms.

3. Speech Recognition models

Once the relevant parameters are selected from the input speech signal, we will use a speech recognition system based on HMM to test our segmentation algorithms. The system is implemented using HTK software tools, with continuous HMM. The speech database used is TI with Connected and Non-Connected Digits .

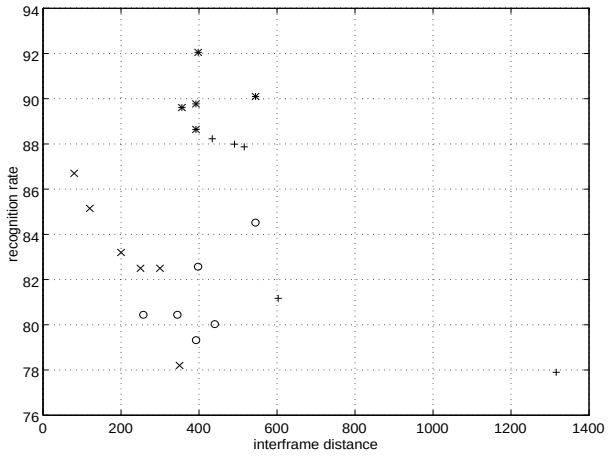

Figure 4: Speech Recognition Rate vs. Interframe Distances, for all the solutions presented in this paper: Uniform Segmentation 'x', Non-Uniform Segmentations based on Maxima Detection and One Gaussian Mixture per Model 'o', Two Gaussian Mixtures per Model '*' and Logarithmic Algorithm '+'

We have selected some alternatives for the models:

Non-Connected Digits In the Reestimation of the models (11 digits + silence) is computed only with Non-Connected sentences, even if the evaluation will take into account both Connected and Non-Connected sentences.

Connected Digits In the Reestimation we take both Connected and Non-Connected Sentences.

In this data base, sometimes (Model *"M5.1"*) we use a different number of states for digits 6 and 8, as they are the shortest and we could have a very few number of frames with some of the solutions tested. We usually work with 8 states per digit, or 8 states for all the digits except digits 6 and 8 that have 6 states. The number of states per digits could not be higher than those values due to the fact that only a few frames are selected per digit (remind the averaged number of frames per digit), compared to other uniformly segmented solutions.

4. Speech Recognition Results

In order to have a good idea of the performance of our algorithms we have tested them with TI data base [9] and their results are shown and commented in this section. In Table 1 the results for a uniformly segmented solution are taken as a reference for the algorithms tested at the Segmentation stage. It is deduced that for higher interframe distances the performance of those uniform models gets worst.

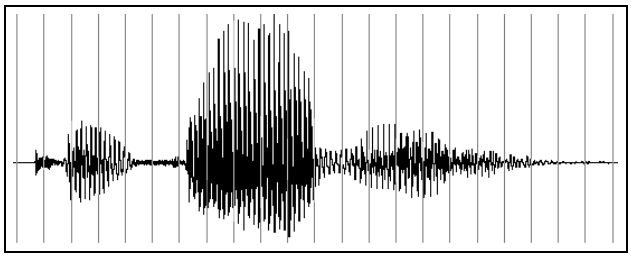

Figure 5: Uniform Segmentation with interframe distance equal to 300 samples for the digits "251" male speaker "aw"

The objective of our algorithms is the improvement of those results with fewer number of frames per digit (or with a higher interframe distance), as a way to prove that in speech recognition systems a good amount of information is redundant and meaningless.

Recognition Rate			
Recognition Rate	*Averaged Interframe Distance*	*Averaged No. of Frames per Digit*	Model Number
80.44	258	14.8	*"M1"*
80.44	345	11.1	*"M2"*
80.03	440	8.7	*"M3"*
79.3	392	9.7	*"M4"*
82.55	398	9.6	*"M5"*
84.5	545	7	*"M6"*

Table 3: Speech Recognition Rate for Maxima Detection Based Non-Uniform Segmentation and one Gaussian Mixture per Model

Working with one Gaussian Mixture per Model, the recognition rate for the Maxima Detection system is better than the uniform one with intyerframe distances higher than 250 samples. Even in the Model *"M6"* the percentage is the highest of all the models, with the highest interframe distance (545 samples or 68.13 msec.). When we work with two Gaussian Mixtures per Model the performance grows 10 points due to a better estimation of the models.

Recognition Rate			
Recognition Rate	*Averaged Interframe Distance*	*Averaged No. of Frames per Digit*	Model Number
89.61	356	10.7	*"M3.1"*
89.77	392	9.7	*"M4"*
88.64	392	9.7	*"M5"*
92.05	398	9.6	*"M5.1"*
90.1	545	7	*"M6"*

Table 4: Speech Recognition Rate for Maxima Detection Based Non-Uniform Segmentation and two Gaussian Mixtures per Model

The solution working with logarithms shows interesting results, due to the important averaged interframe distance (up

to 1319 samples or 164.88msec.) while the recognition rate continues quite acceptable.

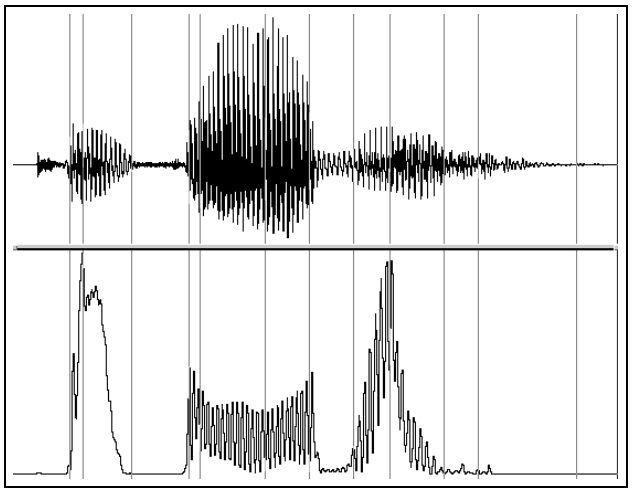

Figure 6: Non-Uniform Segmentation based on maxima location for digits "251" male speaker "aw". Speech Signal, and output from the third band of the MGWTSA

5. Conclusions

In this paper we have shown a new Non-Uniform Segmentation Algorithm working with the time-frequency information delivered by a Modulated Gaussian Wavelet Transform based Speech Analyser, on two of their 17 scales. Doing a pseudo-iterative procedure with the signals in those two scales of the system, we annotate which are the relevant information instants in the speech signal, and build a Continuous Models Speech Recognition system with frames of Time-Frequency information.

Both the Segmentation and the Recognition performance of the system are evaluated, and the paper indicates that working with quite high non-uniform interframe distances the recognition rate of the system will remain acceptable and it could even be improved. Future works on this Segmentation System, will introduce a new algorithm working with the time-frequency information taken from all the scales of the MGWTSA front end.

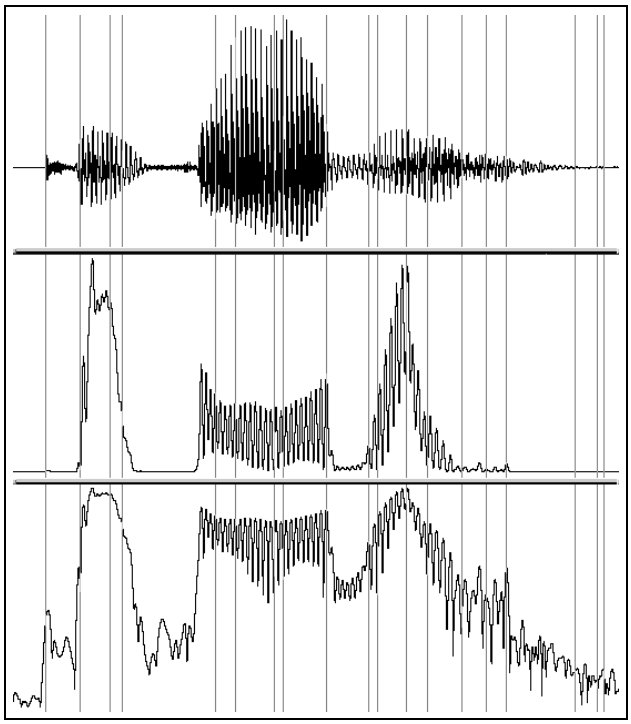

Figure 7: Logarithmic Non-Uniform Segmentation for digits "251" male speaker "aw". Speech Signal, and output from the third band of the MGWTSA, and logarithmic detection signal

6. REFERENCES

1. F.J. Ancin, B.L. Burrows, and R.A. Carrasco. A Novel DyWTVT approach for continuous speech pitch estimation. In *Proceedings EUSIPCO*, volume 3, pages 7P.13 1677–1680, 1994.
2. Mark Black and Mehmet Zeytinoglu. Computationally efficient wavelet packet coding of wide-band stereo audio signals. In *Procedings ICASSP*, volume 5, pages 3075–3078, 1995.
3. F. Cutugno and P. Maturi. Analysing connected speech with wavelets: some Italian data. In *Proceedings EUROSPEECH*, 1993.
4. C. D'Alessandro. Speech Analysis and Synthesis Using an Auditory-Based Wavelet Representation. In *Proceedings ESCA Workshop: Comparing Speech Signal Representations: Sheffield, England*, April 1992.
5. R.F. Favero and F. Gurgen. Using Wavelet Dyadic Grids and Neural Networks for speech recognition. In *Proceedings ICSLP*, 1994.
6. Léonard Janer. Modulated Gaussian Wavelet Transform based Speech Analyser (MGWTSA) Pitch Detection Algorithm (PDA). In *Proceedings EUROSPEECH*, volume 1, pages 401–404, 1995.
7. Won-ha Kim and Yu-Hen Hu. Wavelet packet based optimal subband coder. In *Proceedings ICASSP*, volume 4, pages 2225–2228, 1995.
8. Mikel L. Larreategui, F.J. Ancin, and Rolando A. Carrasco. An Improved Epoch Detection Algorithm Based on Sinusoidal Modelling of Speech. In *Proceedings EUROSPEECH*, volume 1, pages 409–412, 1995.
9. R.G. Leonard. A database for speaker-independent digit recognition. In *Proceedings ICASSP*, pages 42.11.1–4, 1984.