TOWARDS PHONE SEGMENTATION FOR CONCATENATIVE SPEECH SYNTHESIS

Jordi Adell, Antonio Bonafonte

Dept. of Signal Theory and Comunication
TALP Research Center
Technical University of Catalonia (UPC), Barcelona (Spain)
www.talp.upc.es

ABSTRACT

We present a new approach to solve the problem of phone segmentation when preparing databases for concatenative Text-to-Speech synthesis. First, we describe the problem and review the state of the art. Then we present some already existing techniques to perform this segmentation and present our approach based on a Regression Tree to perform Boundary Specific Correction of the HMM segmentation. We discus different evaluation procedures. Finally, we compare some systems and we show how our system improves the system based on HMMs setting 94% of the boundaries within a tolerance of 20ms compared to a manual segmentation, and how phonetic rather than acoustical features are better suited for this task.

1. INTRODUCTION

Nowadays, concatenative speech synthesis is the most widely used approach and it leads the actual performance of Text-to-Speech (TTS) systems. Nevertheless, this approach deals with the problem of needing a large speech database to ensure there is an appropriate unit for the one we are looking for in the selection process. In many situations, the success of the system lays on the correct treatment of the database.

When using concatenative TTS synthesis, we need to spend a big part of the effort on preparing the database. It has to be correctly designed, in order to cover all the variability of the language, and well recorded for the system to have high voice quality. It is important that the database is recorded without noise and it helps if it is a professional speaker. Finally, we need to process the database after recording it, in order to add some information needed by the selection process. Some of this information, such as pitch and energy can be extracted automatically and its cost is very low. However, some other needs, at least at the moment, to be set manually. A couple of these manual high cost tasks are voice segmentation and phonetic labelling.

This work has been partially sponsored by the Spanish Government under grant TIC2002-04447-C02.

Phonetic transcription has different levels of difficulty depending on the language we use. The work we present here was done in Spanish and our system can perform phonetic transcription using a dictionary and rules giving high accuracy [1]. Segmentation is also the most expensive task, that is why we focused on it.

2. PHONE SEGMENTATION

Until now, the highest results have been achieved by manually processing the corpus, but some researchers [2] claim that actual automatic methods for voice segmentation can already achieve accurate enough results for its use in concatenative speech synthesis. They support this claim on perceptual evaluation of the systems. However, the influence of phone segmentation in the final naturalness and intelligibility of the speech depends on the philosophy of each system. It would affect in different manners if we use different units to concatenate. If we need the segmentation of the concatenation points or we look for them automatically the effects would differ. For instance, some systems use diphone segmented and some others phone segmented voices.

However, the present work is focused on phone segmentation, so diphone boundaries have not been considered. Finding diphone boundaries is a task that belongs to the concatenation-point detection framework [3] and boundaries could not be considered to be static but vary from each realization to another.

There already are many different published approaches to this problem. The most studied is the one based on the speech recognition paradigm. Hidden Markov Models (HMM) can be used to perform a recognition task over the voice we want to segment and the edges of HMMs' states will give us the boundaries of the phones [4]. These boundaries can be improved by using some Machine Learning techniques [5].

In the literature we find these and other different approaches to this problem:

- Hidden Markov Models [4, 6]
- Artificial Neural Networks [7]

- Dynamic Time Warping [8, 9]
- Gaussian Mixture Models [7]
- Pronunciation Modeling [10]

In this work we have chosen some of these methods and applied them to the same database. We considered that there is a lack of standard frameworks available to allow comparison between segmentation systems. So we have compared these different approaches by applying them on the same database in the same conditions. This helps in a more objective analysis of the approaches. We also propose a new approach to the problem based on a Regression Tree that achieves good performance.

3. SYSTEMS DESCRIPTION

In this section we will review different methods involved in the present work; their theoretical framework, advantages and disadvantages.

3.1. Hidden Markov Models

This method was one of the first used to attempt to solve the problem presented in this paper [4]. It consists on performing a recognition task over the recorded voice. If we consider that we already know the phone sequence, only an HMM sequence is allowed and models' transitions give us the boundaries of the phones [8, 2].

In the specific case of synthesis using diphones, it could also be necessary to know the diphone boundaries, then we should use demiphone HMMs in order to segment voice into demiphones and then have a mark for phone and diphone boundaries.

3.2. Artificial Neural Networks

Artificial Neural Networks (ANN) can be used to correct the boundaries given by HMM based systems and achieve better performance using this Machine Learning technique [7].

ANN try to estimate the probability of having a boundary in a specific frame from a set of characteristics extracted from the voice. They can either be measured from the signal, such as correlation between frames, pitch, etc. or they can also be qualitative rather than quantitative such as phonetic features.

The network is trained from an already manually segmented corpus, and in test mode it gives as output the probability of a boundary occuring in a specific frame. Then the network should be applied through all the rest of the corpus. The output can be plotted against time as in Figure 1. Then the HMM boundaries (vertical dotted lines) are moved to the closest maximum given by the network (marked with arrows).

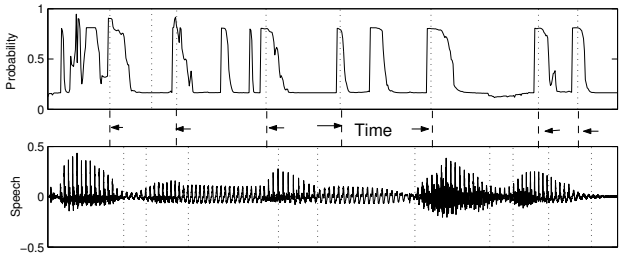

Fig. 1. Estimation of boundary probability using an ANN. Dotted lines show HMM boundaries and arrows its correction.

3.3. Dynamic Time Warping

This method is based on the idea that we can align an already segmented voice with a non-segmented one. A synthesized voice is taken as the already segmented since we can know where the units used to build the voice start and end. So we can map the boundaries of the synthesized voice into the recorded one after aligning them by a Dynamic Time Warping algorithm (DTW).

The dynamic alignment is performed on some characteristics extracted from the signal. Different parameterizations such as Mel-Frequency Cepstral Coefficients (MFCC), Linear Spectral Coefficients (LSF) or Linear Prediction Coefficients (LPC) can be used. They can also be combined in a way that the one that describes better each phonetic class is used for this class [9].

This method implies building a voice for a TTS. In order to do it, we need manually segmented speech and It has to contain all the available units in the language. To solve this it is possible to use an already built voice from another speaker with all the necessary units and voice quality. Another solution would be to gradually segment the database. Starting from a manually segmented part of it, this part can be used to built a new voice. Then, by means of a bootstrap process, we can re-segment the database plus a new part of it and built a new voice from this segmented speech. Iterating this way we can segment the whole database.

3.4. Regression Tree

We propose a new approach to the problem based on the idea of Boundary Specific Correction (BSC)[11]. This approach comes from the idea that HMMs do systematic errors (i.e. The error is similar for similar transitions). This claim is also supported by comments of people that have been manually correcting the HMM segmentation. They comment that HMMs perform better for some transitions

than for others, and for a specific transition they are always mistaken in the same direction.

The method consists on applying a correction of previous calculated boundary (e.g HMM) as a function of some characteristics of the transition. We manually segment part of the voice. Typically a small part of it, probably with a couple of minutes would be enough, but it would also depend on how representative this speech is. Then a Regression Tree is built in order to do a regression of the error between the HMMs boundaries with respect to the manual segmented ones.

In order to do this regression, we used a binary decision tree, and splitting criterias were questions about phonetic properties of phones at left and right boundary sides. This also helps us to correct the boundary in the case that some transition have not been seen in the training data. For this we assume that for similar transition the error is likely to be similar.

After that, we apply this tree to the rest of the voice moving the boundaries by the amount of time given by each leaf of the tree.

4. EXPERIMENTS AND EVALUATION

4.1. Corpus

In order to carry out the experiments we used a corpus recorded in the TALP Research Center. It consists on 516 manually segmented sentences what corresponds to approximately half an hour of speech. It was a female speaker. HMMs were trained on the whole corpus using its phonetic transcription but no segmentation information. For the other methods 40 of those sentences where randomly chosen to become the training corpus. Therefore the results presented are evaluated on the rest of the corpus. For the system based on DTW we used different settings. Since with 40 sentences it gave poor results, we used a set of configurations with 40, 200, 300 and 400 manually segmented sentences.

4.2. Methodology

This paper is oriented to TTS synthesis. Although it would be interesting to use speaker independent models, here, training and testing data always come from the same speaker for all the systems. Therefore, the problem is to generate a new voice for a TTS and if we can have better accuracy despite we must manually segment a small part of the speech, it would help a lot.

HMM. Speaker dependent HMMs for demiphones were used. We used RAMSES, the UPC speech recognition system. The parameterization was MFPC (Mel-Frequency Power Cepstrums), their first and second derivatives and first derivative of the energy. Therefore the parameters were extracted with a 20ms window and a 4ms delay between them. We used semi-continuous HMMs with a codebook of 128 centroids. We first trained the HMMs context independent for 12 iterations and then performed 6 context dependent iterations. The same HMMs results have been used as a starting point for the ANN based and CART based systems. Both try to correct the HMM's boundaries by doing a refinement on them.

ANN. We have used a system based on the one of [5]. A Multilayer Perceptron (MLP) was trained with 9 features extracted from the speech signal. Six of these features were the *energy*, *Zero Cross Rating (ZCR)* and *mean frequency* calculated before and after the boundary. Then two duration features. Both were the value of a Gaussian function starting in the previous boundary and ending in the next one. In one case with its center in the position of the boundary given by the preliminary segmentation, here the HMM. For the second one, the center of the function was a point that satisfied that durations of previous and next phones were proportional to their mean durations evaluated in the manual segmentation. Finally, spectral correlation between the signal before the boundary and the signal after it was the ninth feature.

The MLP had one hidden layer with five units and one single unit in the output layer. It was trained with 1 in the output for boundaries and with some randomly selected frames between the boundaries whose output was set to 0. We used a supervised back-propagation method for training.

CART. Using the 40 sentences of the training corpus a Regression Tree was constructed by asking binary questions about phonetic features (i.e. manner, articulation point, voice, etc...) of boundaries' left and right phones. The tool used was *wagon* from the Edinburgh Speech Tools[12]. We trained a tree to do a regression of the error between the manual segmentation and the one from the HMM based system with minimum 35 units in each leaf. Then the rest of boundaries of the corpus where corrected by using this regression tree, according to phone characteristics at left and right sides of the boundaries.

DTW. We used 40 manually segmented sentences to build up a voice for the UPC-MLTTS, the speech synthesis system from the TALP Research Center [13]. Then, with this voice, we synthesized the rest of the corpus. After that, using this voice, we aligned these synthesized sentences with the recorded ones using the DTW-Festvox utility [14]. The alignment was done with MFCC (Mel-Frequency Cepstral Coefficients) extracted using the Edinburgh Speech Tools. Finally we used this alignment to map the boundaries of the synthesized sentences into the recorded ones.

Results were poor and we noticed an important dependence with the number of sentences used to build up the voice. Thereby, we present three more experiments with different sets of sentences used to build the voice for synthesis.

Finally, in order to compare ANN with CART, we performed two more experiments. They were similar to the ones described above, but using for both systems acoustical and phonetic features. In the case of ANN an integer was assigned to each phonetic feature. We will refer to these experiments as **ANNa** and **CARTa**.

4.3. Evaluation

There are different ways to evaluate the performance of a segmentation system. Evaluation methods can be divided into two different groups: the ones based on an objective measure and the ones based on a subjective ones. For latter, usually perceptual tests are performed. Using the segmented speech we can synthesize some speech and give it to a group of people to evaluate them. For objective evaluation there also are different methods. The most widely used in the literature is to measure the agreement with a manual segmentation. Usually the percentage of boundaries whose error is within a tolerance is measured for a range of tolerances. Another measure proposed in [7] can group tolerances in one single number. It consists on the mean of a set of tolerances that will allow to compare different systems in an easy way. We will refer to this parameter as *MeanTol*.

Objective and perceptual evaluations are complementary rather than contradictory. In [2] they propose a perceptual evaluation for the present task. However, the cost of this evaluation is high since a group of listeners is needed, test needs to be carefully designed and realizations controlled. We have chosen to perform an objective evaluation of the systems.

When doing objective evaluation, some researchers have wondered whether or not a manual segmentation is a valid reference [7, 10]. To evaluate this, they have given the same speech database to different people to segment it. Then, they evaluated the difference between them. In [6] we can find a review of similar experiments, even for different languages. Results reported show an average agreement of 94% within 20ms for human labelers.

These values can be interpreted as a limit for automatic systems when using a manual reference as an objective measure of the accuracy.

5. RESULTS

We calculated the percentage of boundaries within a set of tolerances. These tolerances are 5, 10, 15, 20 and 25 ms. Results are presented in Table 1.

There we can observe how the lowest accuracies correspond to the DTW based system. Also the HMM give low results compared to the results from ANN and CART. These two systems both give similar results and significantly improve the HMMs. Nevertheless, we should notice that the latter systems build on the results of the HMM and try to improve them. We should also remember that HMMs are fully automatic and Machine Learning techniques need a manually segmented corpus as training data.

System	< 5	< 10	< 15	< 20	< 25
HMM	41%	67%	85%	92%	94%
ANN	56%	79%	88%	92%	95%
ANNa	54%	81%	90%	94%	95%
DTW	30%	50%	62%	69%	73%
CART	**58%**	**82%**	**91%**	**94%**	**96%**
CARTa	48%	77%	88%	93%	95%

Table 1. Percentage of boundaries within different tolerances for every system, tolerances are in ms.

Another way for comparing systems is using *MeanTol*. Here we consider this parameter as the mean of the tolerances in Table 1. Values for this parameter are shown in Table 2.

HMM	**ANN**	**ANNa**	**CART**	**CARTa**	**DTW**
76	81	83	84	80	58

Table 2. *MeanTol* for every system.

Using this parameter, we confirm how DTW's accuracy is low. HMM has bigger values although ANN and CART improve them significantly. Therefore, we can observe how CARTa gives lower results than CART. Thus, CART achieves its highest result when using phonetic features on their own. On the other hand, ANN improve its accuracy when adding phonetic features. This clearly shows that phonetic features are better suited for refining HMM boundaries.

In the case of DTW the results are not good. Because of that, we present here some more experiments. They were performed with more manually segmented sentences, and these sentences were chosen, using a greedy algorithm, in order to represent the language variability. The accuracies are shown in Table 3.

DTW Accuracies					
Sentences	< 5	< 10	< 15	< 20	< 25
40	30%	50%	62%	69%	73%
200	37%	61%	72%	80%	85%
300	39%	59%	72%	80%	84%
400	**40%**	**62%**	**77%**	**85%**	**88%**

Table 3. Results for DTW using different sets of manually segmented sentences. Tolerances are in ms.

Although we can observe how the system significantly improves when adding more manually segmented data, it do not reach other systems results. This shows how DTW

would only by a useful system if you could manually segment more training data in order to build a voice for a TTS system. It could also be used to segment speech by using a voice from another speaker built by manually segmented data. On the other hand, ANN and CART based systems are more automatic, since they need less manually segmented resources, about 3 or 5 minutes of speech.

In Figure 2, we have plotted the results for every system, so it is easier to compare them. We have also plotted the manual agreement (MAN) between to people presented in [7]. Note that these values were measured on a different database so it has been plotted *only* to illustrate the limit in the evaluation method. However, referring to this disagreement, we can observe how the systems presented based on Machine Learning techniques are significantly closer to this limit.

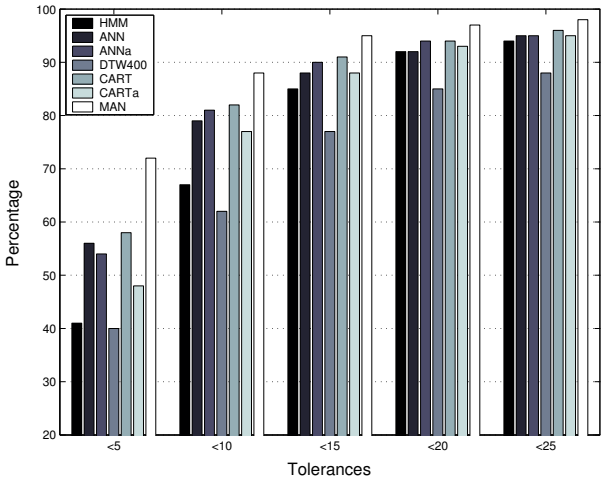

Fig. 2. Graphics of the results for every system.

Apart from these objective measures, we present in Table 4 *MeanTol* parameter for the eight worse transitions for each one of the systems analyzed here. We can see in this table how for HMMs the main problems is with the label CL. This is a label we use to identify the short silence before plosives */p/*,*/t/* and */k/*. HMMs though, cannot segment properly boundaries that come before these silences. On the other hand, both CART and ANN solve this problem, and their main problems turn into voiced phones such as nasals, vowels or laterals. Voiced characteristics increase the most the difficulty of the task.

6. CONCLUSIONS

In this paper we have reviewed the main existing techniques for phone segmentation applied to TTS synthesis. We have also discussed about evaluation methods for this kind of techniques and discussed the limitations of an objective measure based on manual segmented corpus. In order to compare different systems we have performed segmentation on the same database using all of them. It has allowed us to extract some conclusions about different performances of all of them using an objective evaluation.

They have different advantages and disadvantages. DTW cannot achieve same performance as HMM, despite it probably could by using a larger manually segmented corpus. On the other hand, HMMs can perform a fully automatic system, but they are not so good on segmenting the speech, compared with CART and

ANN. Probably because they are designed for recognition without caring about boundaries.

On the other hand, ANN and CART gave the highest results, and in fact they perform a similar regression of the error. They learn the error of the HMMs from a manual segmentation. The advantage for these systems is that they do not need as much training data as DTW and they give high accuracies.

Results presented show how phonetic features can better refine HMMs boundaries than acoustic features. We can observe this because when adding phonetic features to ANN it improves its accuracy and CART gives better results when using phonetic features alone than with all features. That is because HMMs are mistaken in such a systematic way that both CART or ANN can correct it. So we encourage the use of phonetic rather than acoustical features to predict the correction of HMMs boundaries.

As an overview of the present work we claim that under an objective measure, both CART and ANN systems we presented here give the best performance for few training data. Since the CART is simpler and gives higher accuracies, we propose this system to perform phone segmentation. It can handle segmentation with 94% of boundaries closer than 20ms to the manual using only phonetic features.

It is also worth mentioning here that we must notice how the difficulty of segmentation increases when segmenting voiced-to-voiced transitions. Transitions that involve vowels, nasals or voiced fricatives are the most difficult ones in general. Although for every system there are differences these classes are presented for every system in Table 4.

Perceptual test comparing these methods in the framework of the real goal of TTS systems should be performed in future work. Also try to combine different methods (e.g. apply CART to the DTW results). It would also be interesting to find whether this system could or not work under a speaker independent approach, although this it not the typical paradigm. Also the analysis of the errors in order to use this information in the selection process is encouraged.

HMM		ANN		CART		DTW	
aprox. fricative *to* CL	35	voice affricate *to* aprox. fricative	20	nasal *to* voiced plosive	22	lateral *to* nasal	23
nasal *to* nasal	38	nasal *to* voiced plosive	23	voiced fricative *to* trilled	34	voiced plosive *to* semi-vowel	40
voiced affricate *to* CL	40	voiced fricative *to* trilled	37	nasal *to* lateral	41	lateral *to* trilled	40
nasal *to* lateral	41	nasal *to* lateral	43	lateral *to* voiced plosive	49	voiced affricate *to* vowel	45
semi-vowel *to* CL	45	lateral *to* voiced plosive	50	nasal *to* nasal	51	nasal *to* vowel	47
trilled *to* CL	46	nasal *to* nasal	52	semi-vowel *to* lateral	60	nasal *to* trilled	50
lateral *to* CL	50	aprox. fricative *to* unvoiced fricative	53	aprox. fricative *to* voiced affricate	60	nasal *to* aprox. fricative	50
lateral *to* voiced plosive	52	aprox. fricative *to* voiced affricate	53	nasal *to* voiced fricative	65	lateral *to* unvoiced affricate	50

Table 4. *MeanTol* for the four worst transitions for each system. CL stands for pre-plosive silence, it is used before /p/, /t/ and /k/ phones. DTW values are for 400 sentences.

7. ACKNOWLEDGMENT

The authors want to acknowledge David Remacha for the software he developed for its use in the present work.

8. REFERENCES

[1] Albert Febrer, *Síntesi de la parla per concatenació basada en selecció.*, Ph.D. thesis, Technical University of Catalonia (UPC), January 2001.

[2] Matthew J. Makashay, Colin W. Wightman, Ann K. Sydral, and Aliasir Conkie, "Preceptual evaluation of automatic segmentation in Text-to-Speech synthesis," in *Proceedings of ICSLP*, October 2000, Beijin, China.

[3] Esther Klabbers and Raymond Veldhuis, "On the reduction of concatenation artefacts in diphone synthesis," in *Proceedings of ICSLP*, December 1998, pp. 1983–1986, Sydney, Australia.

[4] P.A. Taylor and S.D. Isard, "Automatic phone segmetation," in *Proceedings of Eurospeech*, September 1991, pp. 709–711, Genova, Italy.

[5] D. T. Toledano, A. Hernández Gómez, and Luis Villarrubia Grande, "Automatic phone segmentation," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 6, pp. 617–625, November 2003.

[6] John-Paul Hosom, *Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information*, Ph.D. thesis, Oregon Graduate Institute of Science and Technology, May 2000.

[7] D. T. Toledano, *Segmentación y etiquetado fonéticos automáticos.*, Ph.D. thesis, Universidad Politécnica de Madrid, February 2001.

[8] John Kominek, Christina Bennet, and Alan W. Black, "Evaluating and correcting phoneme segmentation for unit slection synthesis," in *Proceedings of Eurospeech*, September 2003, pp. 313–316, Geneva, Switzerland.

[9] Sérgio Paulo and Luís C. Oliveira, "DTW-based Phonetic alignement using multiple acoustics features," in *Proceedings of Eurospeech*, September 2003, pp. 309–312, Geneva, Switzerland.

[10] Andreas Kipp, Maria Barbare Wesenick, and Florian Schiel, "Pronunciation modeling applied to automatic segmentation of spontaneous speech," in *Proceedings of Eurospeech*, 1997, Rhodes, Greece.

[11] Jindrick Matousek, Daniel Tihelka, and Josef Psutka, "Automatic segmentation for czech concatenative speech synthesis using statistical approach with Boundary-Specific Correction," in *Proceedings of Eurospeech*, September 2003, pp. 301–304, Geneva, Switzerland.

[12] Paul Taylor, Richard Caley, Alan W. Black, and Simon King, "Edinburgh speech tools library system documentation," http://www.cstr.ed.ac.uk/projects/speech_tools/manual-1.2.0/, June 1999.

[13] Antonio Bonafonte, Ignasi Esquerra, Albert Febrer, José A. R. Fonollosa, and Francesc Vallverdú, "The UPC text-to-speech system for Spanish and Catalan," in *Proceedings of ICSLP*, November 1998, Sydney, Australia.

[14] Alan W. Black and Kevin Lenzo, "Building voices in the festival speech synthesis," http://www.festvox.org/bsv.