COMPARATIVE STUDY OF AUTOMATIC PHONE SEGMENTATION METHODS FOR TTS

Jordi Adell, Antonio Bonafonte, Jon Ander Gómez, María José Castro**

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

*Dpto. de Sistemas Informáticos y Computación
Universitat Politècnica de Valencia
Valencia, Spain
www.dsic.upv.es

ABSTRACT

In the present paper we present two novel approaches to phonetic speech segmentation. One based on acoustical clustering plus dynamic time warping and a second one based on a boundary specific correction by means of a decision tree. The use of objective or perceptual evaluations is discussed. Novel approaches clearly outperform objective results of the baseline system based on HMM. They get results similar to agreement between manual segmentations. We show how phonetic features can be successfully used for boundary detection together with HMMs. Finally, the need for perceptual tests in order to evaluate segmentation systems is pointed out.

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. New databases are often needed in order to create new speakers for same language, multilingual purposes, creating a variety of speaking styles or even for adapting a TTS system to a new domain. It is therefore crucial to reduce the amount of effort needed by the process of building these databases.

Parts of this process need, at least at the moment, to be completely manual or manually supervised. Manual tasks demand a large effort, which increases database pre-processing costs and may be inconsistent. Phone segmentation is one of these tasks and automatic segmentation could reduce the effort requested.

In order to attempt the problem of automatic phone segmentation, we can choose among three different approaches depending on the previous information we may have: *unconstrained*, *acoustically constrained* or *linguistically constrained* [1]. In the present paper a linguistically constrained approach have been considered. Automatic phone transcription is a problem we are not taking into account here, thus we assume we already know the correct transcription of the database (e.g. by means of a lexicon plus a manual correction).

Although some researchers claim that actual automatic segmentation systems can already achieve accurate enough results for their use on speech synthesis [2], many research labs still get their best results by manually supervising the data. Therefore, in the literature we can still find many research works on a variety of methods such as: *Hidden Markov Models* [3, 2], *Artificial Neural Networks* [4] or *Dynamic Time Warping* [5, 6].

There is a lack of standard frameworks available to allow comparison between segmentation systems and there is no agreement in the literature on how segmentation must be evaluated. This is why in the present paper evaluation methods are discussed. In this work we have chosen some methods and applied them to the same database in the same conditions. Furthermore, we propose two novel methods to approach phone segmentation problem. In next section we describe the chosen methods, in Section 3 we discuss their evaluation and finally results and conclusions are presented.

2. METHODS DESCRIPTION

In this section we will review different methods involved in the present work, their theoretical framework, advantages and disadvantages. Methods 2.1 and 2.2 are the most used in the literature. We propose two new methods to overcome their limitations.

2.1. Hidden Markov Models

This method was one of the first used to attempt to solve the problem presented in this paper [3]. It consists on performing, since we know the phonetic transcription, a forced alignment by means of the Viterbi algorithm. Transition between models are then considered as phone boundaries.

We used RAMSES, the UPC speech recognition system. Speaker dependent HMM-demiphone models were used [7]. Parameterization was MFPC (Mel-Frequency Power Cepstrums), their first and second derivatives and first derivative of the energy. Parameters were extracted with a 20ms window and a 4ms delay between frames. We used semi-continuous HMMs with a codebook of 128 Gaussians. We first trained the models context independent for 12 iterations and then performed 6 context dependent iterations.

This work has been partially sponsored by the European Union under grant FP6-506738 (TC-STAR project, http://www.tc-star.org) and the Spanish Government under grant TIC2002-04447-C02 (ALIADO project, http://gps-tsc.upc.es/veu/aliado).

2.2. Acoustic alignment by Dynamic Time Warping

Dynamic Time Warping (DTW) together with a synthesized voice has been widely used since its first attempt [6]. This method uses a dynamic algorithm to align an already segmented voice with a non-segmented one. Synthesized speech is aligned with a recorded one which has the same phones. In TTS the database is labeled so we know where the phone boundaries are in the synthesized speech. Then, these boundaries are mapped onto the recorded speech by means of the alignment performed.

A manually segmented sub-corpus was used in order to build up a voice for the UPC-MLTTS, the speech synthesis system from the TALP Research Center [8]. Then, using this voice, we synthesized the rest of the corpus. After that, we aligned these synthesized sentences with the recorded ones by means of the DTW-Festvox utility. MFCC (Mel-Frequency Cepstral Coefficients) were used for mapping and extracted using the Edinburgh Speech Tools.

2.3. Acoustic Clustering-Dynamic Time Warping

In this subsection we propose the Acoustic Clustering-Dynamic Time Warping speech segmentation technique (AC-DTW), an approach to automatic speech segmentation based on unsupervised learning of acoustic classes and its association to phonemes by means of conditional probabilities.

Phonetic boundaries are established by a Dynamic Time Warping algorithm that uses the *a posteriori* probabilities of each phonetic unit given an acoustic frame. These *a posteriori* probabilities are calculated by combining probabilities of acoustic classes, which are obtained from a clustering procedure on the acoustic feature space, and the conditional probabilities of each acoustic class with respect to each phonetic unit [9].

In the clustering procedure, it is assumed that acoustic classes can be modeled by means of Gaussian distributions. Parameters of each Gaussian distribution are estimated by using the unsupervised version of the Maximum Likelihood Estimation procedure [10]. Thus, it is possible to estimate the probability of each acoustic class w_c given an acoustic vector x_t, $\Pr(w_c|x_t)$, from the Gaussian Mixture Model. Nevertheless, as we need the probability of each phonetic unit ph_f given an acoustic vector x_t, $\Pr(ph_f|x_t)$, a set of conditional probabilities are estimated in order to calculate the phonetic probabilities from the acoustic ones.

The use of conditional probabilities allows us to compute the phonetic-conditional probability densities $p(x_t|ph_f)$ as follows [9]:

$$p(x_t|ph_f) = \sum_{c=1}^{C} p(x_t|w_c) \cdot \Pr(w_c|ph_f) \qquad (1)$$

where C is the number of acoustic classes, $p(x_t|w_c)$ is the acoustic class-conditional probability density, computed as the Gaussian probability density function, and $\Pr(w_c|ph_f)$ is the conditional probability that acoustic class w_c has been manifested when phonetic unit ph_f has been uttered.

Then, applying the Bayes formulation we obtain the phonetic probabilities as:

$$\Pr(ph_f|x_t) = \frac{\sum_{c=1}^{C} p(x_t|w_c) \cdot \Pr(w_c|ph_f)}{\sum_{j=1}^{F} \left(\sum_{c=1}^{C} p(x_t|w_c) \cdot \Pr(w_c|ph_j) \right)} \qquad (2)$$

where F is the number of phonetic units. Finally, a DTW algorithm is used in order to align the frame sequence with the phonetic transcription by means of conditional probabilities mentioned above.

For this method, the set of conditional probabilities is computed from a sub-corpus of sentences manually segmented and labeled. Each acoustic frame x_t is formed by a d-dimensional vector: the normalized energy, the first 12 MFCC and their first and second time derivatives. An acoustic frame is obtained every 4 ms. using a 20 ms. Hamming window.

2.4. Regression Tree-based Boundary Specific Correction

We present a new approach based on Boundary Specific Correction (BSC) [11]. The presented approach has two steps. In the first step a coarse segmentation is performed. The second one consists on refining these boundaries. This *two steps* technique has already been used as in [12].

Typically, acoustic-based approaches have been considered in the literature [13]. Acoustic approaches claim that boundaries can be detected by measuring local acoustic dynamics. However, our previous experiments have shown that phonetic features are better suitted for refining HMM's boundaries than acoustic features [14]. Therefore, in our approach, we introduce, in the second step, phonetic information. Boundaries are refined using the phonetic features (i.e. manner, articulation point, voice, etc...) of both phones involved in the transition. Thus, in the second step we propose to use a Regression Tree (RT). A small sub-corpus is used to train a RT that makes a regression of the error between the manually supervised and the HMM-based segmentation as a function of the phonetic features. Then, this tree can predict the error for the rest of the corpus, thus it can be corrected.

This is supported on the idea that HMMs perform similar errors for phonetically similar transitions. This 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 that for a specific transition they are always mistaken in the same direction.

In the present work 40 sentences (i.e. a couple of minutes of voice) have been used in order to train the tree. The tree was built by using binary questions about the phonetic features of the boundary context. After that, this tree is applied to the rest of the voice moving each boundary the amount of time given by the corresponding leaf of the tree. The use of decision trees additionally helps us to correct boundaries in the case that some transitions have not been seen in the training data.

The Regression Tree was constructed using a training corpus and the tool *wagon* from the Edinburgh Speech Tools[15]. We trained a tree with minimum 35 units in each leaf.

3. SEGMENTATION EVALUATION

The evaluation criteria most widely used in the literature is to measure the agreement with respect to a manual segmentation. Usually the percentage of boundaries whose error is within a tolerance is calculated for a range of tolerances. In [4] it is also proposed to calculate the mean of values for a range of tolerances, hence one single value is obtained. This allows an easy comparison between systems. We will refer to this parameter as *MeanTol*.

When doing this objective evaluation, some researchers have wondered whether or not a manual segmentation is a valid reference [4, 16]. To evaluate it, they have given the same speech

database to different experts to segment it. Then, they evaluated the difference between them. As a result in [4] they obtained 97% of the boundaries within a tolerance of 20ms and in [16] 93%. We interpret this agreement as the maximum accuracy for a segmentation system, since a system that reaches 100% compared with a manual segmentation will at least differ around 95% with another one for the same speech database.

On the other hand, in [2] they propose a perceptual evaluation in order to compare segmentation systems. A perceptual evaluation can measure the real goal of the application and allows us to know whether segmentation differences have a real influence in the final goal. As a result, we propose objective measures, since their cost is lower, for a first comparison. But for final comparison when objective accuracies are high, a perceptual test would help us to discuss whether a new segmentation is worth using for TTS systems.

4. EXPERIMENTAL RESULTS

4.1. Corpus

In order to carry on the experiments we used a corpus recorded in the TALP Research Center. It consists on 516 manually segmented sentences, what means about half an hour of speech. It is a female speaker and style is neutral. 40 of these sentences where randomly chosen to become the training corpus for AC-DTW and RT-BSC and used to create a voice for the DTW method. Therefore, results presented are evaluated on the rest of the corpus.

4.2. Objective Test

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.

System	< 5	< 10	< 15	< 20	< 25	*MT*
HMM	41%	67%	85%	92%	94%	76
DTW	30%	50%	62%	69%	73%	58
AC-DTW	**52%**	**78%**	**89%**	**93%**	**96%**	**82**
RT-BSC	**58%**	**82%**	**91%**	**94%**	**96%**	**84**

Table 1. Percentage of boundaries within different tolerances for every system (*tolerances in ms*). Also the *MeanTool* (MT) value for every system is presented

In Table 1 we can observe how the lowest accuracies correspond to the DTW-based system and both AC-DTW and RT-BSC highly improve HMMs results.

Dynamic Time Warping algorithm is considered to be more precise than HMMs in mean, while its problem is that errors can be very large [5]. In Table 2 it can be observed how when only considering errors smaller than 20ms HMM still improves DTW accuracies. Then, we cannot consider DTW to be more precise than HMMs in any sense.

Systems	< 5	< 10	< 15	MT
DTW	44%	74%	90%	69
HMM	44%	77%	93%	**71**

Table 2. Results considering only errors lower than 20ms.

We also present here some more experiments about using DTW algorithm. They were performed with more manually segmented sentences to build up the voice for synthesis, and these sentences were chosen using a greedy algorithm in order to represent the language variability. Accuracies are shown in Table 3.

DTW Accuracies						
Sentences	< 5	< 10	< 15	< 20	< 25	MT
40	30%	50%	62%	69%	73%	57
200	37%	61%	72%	80%	85%	67
300	39%	59%	72%	80%	84%	67
400	**40%**	**62%**	**77%**	**85%**	**88%**	**70**

Table 3. Results for DTW using different sets of manually segmented sentences (*tolerances in ms*).

We can observe how the system significantly improves when adding more manually segmented data. However, even using 400 sentences results do not reach the other methods. These observations discard this method, as used here, for automatic segmentation. However, the two novel approaches presented got similar results as human agreement mentioned in Section 3.

4.3. Perceptual Test

Since we observed that when using few manually segmented data, DTW algorithm does not reach an appropriate performance, we have removed it from the perceptual test. Only HMM, AC-DTW, RT-BSC and Manual Segmentation have been tested.

For this test, 50 different sentences have been synthesized using each segmentation system and the UPC-MLTTS [8], which is a unit selection TTS system and TD-PSOLA was used only for units that differed more than 15ms or 20Hz from the target. The 476 sentences of the test database were used to build the units catalog and prosody was extracted from natural speech in order to avoid effects produced by the prosody model. Sentences have been presented to each of a group of 10 participants, as a set of 20 sentences chosen randomly among the 50. Each sentence was synthesized using two different systems (also chosen randomly) and participants were asked to say which of them was more natural. They could mark: *equal*, *more natural* or *much more natural*. This allowed us to compare each system against each other.

Results of the test are presented in Figure 1 and 2. There we can see how HMMs have always been preferred to any other system and equally preferred with manual segmentation. RT-BSC is only preferred to AC-DTW.

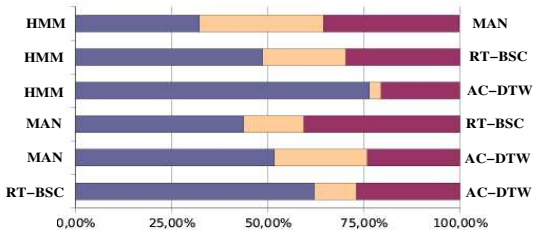

Fig. 1. Percentage of times a system was preferred to each other. Dark colours show percentage of times a systems in edges were preferred and light colours show percentage of times they were considered equally natural.

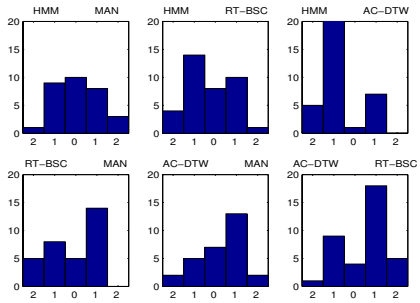

Fig. 2. Answers distributions for each pair of systems. Numbers in horizontal axis mean: 0-*equal*, 1-*more natural* and 2-*much more natural*. Names on top of figures show systems being compared.

In Figure 2 distributions of the answers are presented. HMM vs. MAN present a flat distribution, what shows that they are comparable. Other distributions show a clear bias to HMM or MAN systems.

When analyzing perceptual results it must be taken into account that for comparing four systems, as here is done, a large amount of participants would have been desired. Nevertheless, in order to overcome this, consistency through participants have been checked and answers appeared to be consistent with each other participants. Then, results from perceptual test can then be trusted.

Results reached by HMM-based system are close to the given by the manual segmentation. Nevertheless, manually segmented data is not expected to have large errors while HMM are. This is shown in Figure 2 where the number of times a MAN sentence is marked *much more natural* is bigger than an HMM is. This points out that outlier elimination (i.e. automatic removal of undesired units) [17] could be useful for improving databases.

Novel systems presented here reached a large improvement in objective measures as shown in Table 1. This improvement was not converted into a perceptual improvement of the voice quality.

5. CONCLUSIONS

In the present work, we have provided a framework that covered the lack of evaluation frameworks for segmentation systems.

We have presented two novel approaches that outperform the baseline system based on HMM. Results show it is possible to reach human level accuracies by simply refining HMM-based segmentation by means of a Regression Tree based on phonetic boundaries to perform Boundary Specific Correction.

Most of the methods presented in the literature have been typically compared to manual segmentations. Here we have shown that this is not enough to warranty a real improvement of the system's performance. Therefore, we recommend, as proposed by [2], the use of perceptual tests in order to evaluate if new segmentation systems have a real influence in final performances.

6. REFERENCES

[1] A. Marzal and E. Vidal, "A review and new approaches for automatic segmentation of speech signals," in *Proceedings EUSIPCO, Barcelona*, 1990, pp. 43–53.

[2] M. J. Makashay, C. W. Wightman, A. K. Syrdal, and A. Conkie, "Preceptual evaluation of automatic segmentation in Text-to-Speech synthesis," in *Proceedings of ICSLP*, October 2000, beijin, China.

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

[4] D. T. Toledano, "Segmentación y etiquetado fonéticos automáticos." Ph.D. dissertation, Universidad Politécnica de Madrid, February 2001.

[5] J. Kominek, C. Bennet, and A. W. Black, "Evaluating and correcting phoneme segmentation for unit slection synthesis," in *Proceedings of Eurospeech*, September 2003, pp. 313–316, geneva, Switzerland.

[6] F. Malfrère and T. Dutoit, "High quality speech synthesis for phonetic speech segmentation," in *Proceedings of the European Conference On Speech Communication and Technology*, 1997, pp. 2631–2634, rhodes, Greece.

[7] J. B. Mariño, A. Nogueiras, P. Pachès, and A. Bonafonte, "The demiphone: an efficient contextual subword unit for continuous speech recognition." *Speech Comunication*, vol. 32, no. 3, pp. 187–197, October 2000.

[8] A. Bonafonte, I. Esquerra, A. Febrer, J. A. R. Fonollosa, and F. Vallverdú, "The UPC text-to-speech system for Spanish and Catalan," in *Proceedings of ICSLP*, November 1998, sydney, Australia.

[9] J. Gómez and M. Castro, "Automatic Segmentation of Speech at the Phonetic Level," in *Structural, Syntactic, and Statistical Pattern Recognition*, T. C. et al., Ed. Springer-Verlag, 2002, vol. 2396, pp. 672–680.

[10] R. O. Duda, P. E. Hart, and D. G. Stork., *Pattern Classification*, 2nd ed. John Wiley and Sons, 2001.

[11] J. Matousek, D. Tihelka, and J. 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] D. T. Toledano, A. H. Gómez, and L. V. Grande, "Automatic phone segmentation," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 6, pp. 617–625, November 2003.

[13] A. Bonafonte, A. Nogueiras, and A. Rodríguez-Garrido, "Explicit segmentation of speech using gaussian models," in *ICSLP*, 1996.

[14] J. Adell and A. Bonafonte, "Towards phone segmentation for concatenative TTS," in *Proceedings of the 5th ISCA Workshop on Speech Synthesis*, July 2004, pp. 139–144, Pittsburgh, Pennsylvania.

[15] P. Taylor, R. Caley, A. W. Black, and S. King, "Edinburgh speech tools library system documentation," http://www.cstr.ed.ac.uk/projects/speech_tools/manual-1.2.0/, June 1999.

[16] A. Kipp, M. B. Wesenick, and F. Schiel, "Pronunciation modeling applied to automatic segmentation of spontaneous speech," in *Proceedings of Eurospeech*, 1997, rhodes, Greece.

[17] J. Kominek and A. W. Black, "Impact of durational outlier removal from unit selection catalogs," in *Proceedings of the 5th ISCA Workshop on Speech Synthesis*, July 2004, pp. 155–160, Pittsburgh, Pennsylvania.