CONCATENATIVE SPEECH SYNTHESIS FOR EUROPEAN PORTUGUESE

Pedro M. Carvalho[i]*, Luís C. Oliveira, Isabel M. Trancoso, M. Céu Viana*,*

INESC/IST, *CLUL

INESC, Rua Alves Redol, 9, 1000 Lisboa, PORTUGAL

{Pedro.Carvalho, Luis.Oliveira, Isabel.Trancoso, mcv}@inesc.pt

ABSTRACT

This paper describes our on-going work in the area of text-to-speech synthesis, specifically on concatenative techniques. Our preliminary work consisted in investigating the current trends in concatenative synthesis and the problems that could arise when we apply the existing state-of-the art solutions to the specific case of European Portuguese.

Our ultimate goal is to develop a text-to-speech system that could be trained for any speaker's voice in a fully automatic way, i.e., we would like to develop a customized text-to-speech synthesizer for any voice reading a predetermined text.

Our first steps in this direction involved such issues as automatic segmentation and alignment of recorded speech, optimized inventory design for concatenative synthesis, unit selection and optimal coupling of the selected units.

1. INTRODUTION

This paper presents our latest progress concerning text-to-speech synthesis in European Portuguese. The joint effort of the two complementary teams (linguists and engineers) involved in this project started in the beginning of this decade with the development of a rule-based formant synthesizer (DIXI) [1]. Several versions of this synthesizer were implemented in the following years, namely to cope with the needs of the handicap community in Portugal.

In parallel with the development of these special-purpose applications, we have been investing in different synthesis models based on concatenative techniques. This includes not only the development of classic PSOLA diphone-based techniques [12], but also the development of CHATR-like systems [5][11], where larger units are selected and concatenated based on prosodic criteria.

Concatenative text-to-speech systems can, in theory, produce very naturally sounding synthetic speech, since they simply join pre-recorded segments or units to form any sentence. In practice, several factors contribute for less perfect speech output quality. For instance, the choice of the best set of pre-recorded speech units that can be used as building blocks is a difficult task. Moreover, the concatenation of units recorded using different intonation or phonetic contexts may produce sub-optimal results even if the set is reasonably complete and if some prosodic transformations are performed during the concatenation phase. Time domain discontinuities and spectral mismatch may also arise and need to be dealt with in the concatenation process.

We have tried to address these problems in the context of the development of a customized text-to-speech synthesizer, i.e., a system that could be trained in a fully automatic way for any user's voice reading a predetermined text. The fully automatic restriction implies that some tradeoffs must be accepted namely in what concerns the construction of an inventory of acoustic units and the determination of the optimal coupling of inventory units.

The investment in terms of concatenative based speech synthesis generally begins with the design and recording of a high quality corpus, in a controlled environment. The manual transcription and alignment of large corpora is extremely time-consuming and requires a profound knowledge of phonetics to accurately time align the transcription labels. Therefore, automatic segmentation / alignment systems are usually adopted to speed up this procedure. Inventory building, however, generally implies labeling the cut points which correspond to optimal coupling of inventory units, a procedure which also needs to be done in a fully automatic way to comply with our constraints.

This paper will try to explain our preliminary work for rapid deployment of a concatenative synthesizer using these constraints. This will be done in four additional sections: the first one describes the corpus used as a basis for this work; the second section discusses the corpus segmentation and alignment. The problems of unit concatenation are described in section 4, with a particular emphasis on the determination of optimal cut points. The next section includes a brief description of our implementation of both the diphone-based concatenative speech synthesizer and the variable length concatenative synthesizer. The latter, in particular, is still in its earliest stages of development. Hence most of the future work described in the last section is devoted to this synthesizer.

2. CORPORA

Two spoken corpora were used in this work: the first one, EUROM.1 [6], has been recorded in the scope of the European project SAM_A (Speech Technology Assessment in Multilingual Applications); the second one, BDFALA [7], has been recorded in a national project with the same name, whose purpose was primarily to extend the core database created in the previous project, primarily for speech synthesis research purposes. Both corpora are summarized in [2].

Only a subset of the EUROM.1 was manually transcribed and time-aligned by expert phoneticians - the *few talkers* subset. Since our first step in this work was to design a speaker-dependent aligner, we have just used for training the material from this subset spoken by one male and one female speakers, consisting of 15 passages of 5 sentences each. For testing, we have used 5 additional filler sentences and 2 extra passages from each speaker. All together, the material from each speaker amounts to around 4000 PLUs (*Phone Like Units*) in the training set and 950 in the test set.

From the BDFALA corpus, we have also used a subset spoken by the same two speakers, which includes a large inventory of logathomes, sentences and isolated words. The subset of logathomes is particularly relevant to this work. It includes about 3100 diphones, in order to cope with the relative importance of stressed position in European Portuguese. Due to the relative large number of different diphthongs, these were not regarded as basic units for concatenation, except in the case of nasal diphthongs, which were much fewer.

Although the two corpora were not recorded in quite the same acoustic conditions (anechoic chamber and sound-proof room, respectively), our original plan was to use the EUROM.1 material to develop tools that later would be applied (and enhanced) using the much larger BDFALA corpus.

3. AUTOMATIC SEGMENTATION AND ALIGNMENT

The design of a fully automatic speaker-dependent alignment system for European Portuguese was done in two stages. In the first one, we built an aligner based on the phonetically transcribed material of EUROM.1. In the second stage, we used a similar method to train an aligner based on the subset of logathomes of the BDFALA corpus. Since the narrow phonetic transcription required manual intervention, the first aligner is basically intended as a reference tool, whose results are to be compared with the ones of an aligner built without manual intervention.

The core of both alignment tools is a speaker dependent HMM monophonic network consisting of 60 PLU models. The models include occlusive and burst parts of stop consonants and stressed and unstressed variants for vowels. Each model is a classic three-state left-to-right model with no skips between the states and three Gaussian mixtures, except for the silence model that has five states. The input vector for the HMM is composed of 12 Mel frequency cepstral coefficients, normalized energy, and their respective first and second order delta coefficients. The input vector was computed every 5ms using a 25ms Hamming window. The system was implemented using the HTK toolkit from Entropic Cambridge Research Laboratory.

Special care was taken in training the EUROM.1 aligner since both training and test sets were very small. We devised a two-phase process to train this alignment tool [3]. The first phase includes the initialization of PLU models, followed by re-estimation and successive iterations of embedded re-estimation and Viterbi alignment until a stop criterion is met, i.e., until the maximum absolute difference between the time-aligned labels produced in two successive iterations is less or equal to the input vector rate (5 ms). The second phase is very similar, with successive iterations of re-estimation and embedded re-estimation as well. Here, however, each iteration is performed using the output time-labels of the previous iteration, instead of the manual labels used in the first phase. The result of the first stage is therefore tuned to produce the best results on the training set and the second stage “flattens” the decision areas of the HMM models in order to try to stabilize the alignment results.

For each PLU transition, we computed the maximum positive, maximum negative, average, absolute and RMS differences between the manual and automatic time-aligned label files, and verified the need for a de-biasing process. The tool yielded in 90% of the segments of the test set an error below 22 ms, for the two speakers of our test set (see Table 1). Worst results were achieved in vowel-vowel, glide-vowels and nasal vowel-glide transitions (as referred by other authors for different languages [13]).

The results reinforced our idea that the HMM models converged into some unknown features from the input vectors that obviously differs from the expert phonetician criteria for manual alignment. This was partially compensated using a de-biasing process based on the average PLU transition alignment error matrix. An improved accuracy of 5 ms in 90% of the cases on the test set was achieved with de-biasing.

It is worth mentioning at this point that the narrow phonetic transcription of the EUROM.1 corpus was produced in a semi-automatic way, using a speaker-independent Viterbi aligner as well to produce initial time labels which were then manually corrected. The re-trained HMM models were then used in a bootstrap process to align a larger amount of data [2].

Speaker	RMS	<10ms	<20ms	<30ms	90%
Female	0.45 ms	74%	89%	96%	21 ms
Male	0.44 ms	73%	90%	95%	20 ms

Table 1: Performance scores for each speaker (after de-biasing) for the EUROM.1 based alignment tool.

The training process of the second aligner we have implemented was fully automatic, not requiring manually aligned time-labels at any stage. As mentioned before, the training data for this aligner was the logathomes subset of the BDFALA corpus. The training is very similar to the second stage of the EUROM.1 aligner, except that it now uses 64 PLU models to account for nasal diphthongs since we now have sufficient training material to deal with them. The alignment results on the same test set of EUROM.1 are presented in Table 2. Worst cases occur in the same transitions as before, plus vowel-glide and voiced-unvoiced stops transitions.

Speaker	RMS	<10ms	<20ms	<30ms	90%
Female	1.10 ms	46%	70%	83%	42 ms

Table 2: Performance scores for the female speaker for the BDFALA-based alignment tool in EUROM.1 test set.

The accuracy of the BDFALA based aligner is about half of the one of the EUROM.1 based aligner, which is not surprising given the fact that the models are trained from scratch. We hope to improve this accuracy by training our BDFALA aligner using initial speaker-independent models trained on the full EUROM.1 labeled corpus.

It could be argued if a more accurate aligner is needed to segment databases for use in a concatenative synthesizer environment. Our hope is that the inventory selection (and/or concatenative algorithms) can cope with this lack of accuracy when computing an optimal cut point for each pair of inventory units. That is, the alignment marks can be just used as rough indicators for the transition areas between PLUs to determine a precise “stable-area” cut point. This is specially true in vowel based transitions (one of the worst case results for the aligner). The determination of the cut point is discussed in the following section.

For the PSOLA-based diphone concatenation, it is important to obtain pitch epochs. These were automatically computed on the basis of the LPC residual, and later smoothed in the regions where the output of the epoch detector strongly differs from the estimated fundamental frequency. The last step in the constitution of our diphone inventory consists of automatically correcting the segment boundaries determined before by the nearest pitch epoch. At this stage, we observed that the voicing decision (a sub-product of pitch synchronous analysis) could be used to further tune the alignment, although no changes were implemented at this time.

4. UNIT CONCATENATION

Having a rough location of the segment’s boundaries, the next step is to devise a strategy to locate the optimal cut points for unit concatenation. We have started this study in the scope of the concatenation of diphones excerpted from the subset of logathomes. Much of this discussion, however, can hopefully be applied to corpora designed for the concatenation of larger units.

Once the spoken corpus has been segmented, fairly simple rules can be designed to place the diphone boundaries in the subset of logathomes. For instance: place the cut point in fricatives in the mid point of the segment boundaries; for vowels, place it in the pitch epoch closest to the mid point; for plosives, use the boundary between closure and burst as the cut point, etc.. This approach is rather dependent on the accuracy of segment boundary determination and, therefore, specially error-prone when this segmentation is done automatically, as we desire. Moreover, even if the mid point is accurately determined, this does not guarantee that the spectral mismatch between the units to be concatenated will be minimized.

This led us to consider an alternative approach based on the use of a spectral mismatch or distortion measure [8][10]. The idea is to concatenate two units in the point which minimizes the spectral distortion between them. Two distortion measures where informally tested based on Mel frequency cepstral coefficients (MFCC), appended with energy and first order differences: pitch synchronously cepstral analysis and frame synchronously (i.e., with a 5 ms period). In both cases, the Euclidean distance was used as a distortion measurement.

For the comparison of the performance of the spectral measures, 10 diphones were randomly selected from each PLU, and optimal cut points generated for each pair.

Figure 1 represents the histograms of spectral distortion values obtained with the frame synchronously method and, for reference sake, with the simpler mid-point based approach described before. The concatenation spectral mismatch must be related with the spectral variations in the neighborhood of the cut point. The spectral distortion values were then normalized by the spectral distance of the two consecutive frames on the cut point. As expected the optimized cut point has a lower average distortion (1.62) than the of mid point distortion (3.5). The optimized average is within the same range as the frame-to-frame spectral discontinuities of each unit

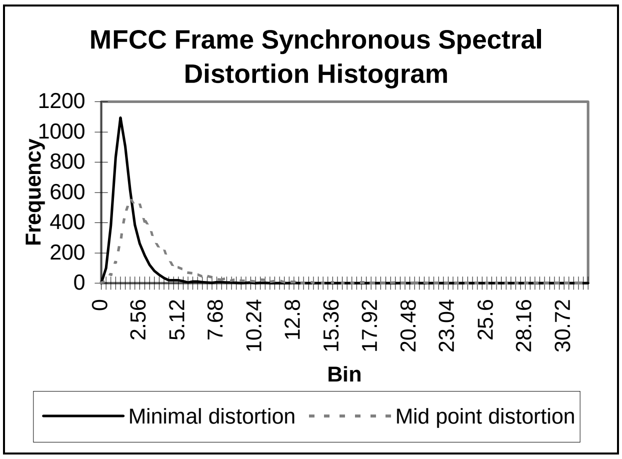

Figure 1: Histogram of spectral distortion values.

The pitch synchronous based distortion measure yielded similar results and therefore was discarded since it involves additional computational effort.

The computational burden of the determination of the spectral distortion from all possible pairs of cut points (pitch epoch marks or 5 ms frame marks) must be considered. This can either be done at run time (synthesis time), which was our approach for the time being, or pre-computed and stored.

5. CONCATENATIVE SYNTHESIS

5.1 Diphone Concatenation

For a rapid implementation and evaluation of a European Portuguese diphone-based concatenative synthesizer, we replaced the formant synthesizer module of the DIXI system by our own implementation of a basic TD-PSOLA [4] synthesis module.

The first informal listening tests allowed us to identify the main problems of the automatic creation of diphone inventories: errors in phone alignment and prosodic marking.

Although the initial set of logathomes was built in order to avoid the effects of the shortening and deletion of unstressed vowels, informal listening tests pointed out this segments as the most problematic. Analyzing this problem, spectral mismatches were found in the synthesized signal: the cut point located by the algorithm was not good enough due to the rapid spectral variations and the small size of the segment.

At this point is safe to say that an insufficient number of tests was performed. As part of our ongoing work we will analyze these issues in depth, specifically the optimal coupling of the diphones and the development of a more formal listening test procedure.

5.2 Variable Length Unit Concatenation

To cope with European Portuguese strong intra and inter-syllabic coarticulatory effects, the concatenative synthesizer environment we developed is prepared to use variable length units (also known as non-uniform units). Diphones constitute, thus, a particular case (length=2) and there are no restrictions for units larger than that.

The basic diphone inventory is being augmented with some typical consonant clusters and other larger units comprising reduced vowels. These units are currently extracted from the subsets of sentences and isolated words of the BDFALA corpus spoken by the same two speakers: 600 phonetically richer sentences and around 4000 words.

In spite of the large number of problems which have still to be dealt with, informal listening tests of synthetic speech produced by the current rough version of this synthesizer yielded encouraging results.

Although the BDFALA corpus was designed to cover several cases of European Portuguese vowel diphthonguisation, coalescence and deletion, as well as of consonant lenition, only a small part of these materials are fully treated. In order to pursue this approach, a much larger set of adequately labeled speech materials is needed for the training and testing of unit selection algorithms for this language. Meanwhile, we plan to cope with this problem by using the augmented inventory and by adding prosodic information to the unit selection algorithm. Drawing on [5], we are aiming at an inventory with several candidates to the same unit in different prosodic contexts. This will also allow us to discard large prosodic changes in the concatenation process.

Several algorithms, like the one described in [9] were already suggested to deal with the problems posed by multi-candidate variable unit length concatenative systems. Our next step will be the implementation of a variable unit selection algorithm for European Portuguese.

6. CONCLUSIONS

In order to speed up the deployment of a concatenative synthesis environment for European Portuguese some important issues were left to be further investigated latter. For instance the alignment tool developed could be refined by training the large BDFALA material with initial time aligned labels produced by a speaker independent EUROM.1 aligner.

For the moment we are using the EUROM.1 alignment tool which yield about 22 ms of precision in 90% of the cases, to segment the BDFALA subset of logathomes. This subset was used to construct a concatenative diphone synthesizer using an MFCC based spectral distortion measure to determine the optimal cut points.

Our current work is focused in two main areas: augmenting the diphone inventory by larger units in order to cope with consonant clusters and vowel reduction effects typical of European Portuguese, and implementing a unit selection algorithm.

We hope that the concatenative synthesis environment that we described in this paper will help creating an ideal investigation tool in order to fulfill our goal of developing an high quality concatenative synthesis environment for European Portuguese.

7. REFERENCES

1. Oliveira, L.C., Viana, M.C., and Trancoso, I.M. - "A rule-based text-to-speech system for Portuguese", *In Proc. Int. Conf. on Acoustic Speech and Signal Proc.*, volume 2, pages 73-76, San Francisco, March, 1992

2. Martins, C., Mascarenhas, M.Isabel, Meinedo, H., Neto, J.P., Oliveira, L.C., Ribeiro, C., Trancoso, I.M., and Viana, M.C., "Spoken Language Corpora for Speech Recognition and Synthesis in European Portuguese, *Proc. of the 10th Conference on Pattern Recognition, RECPAD'98*, pages 357-364, Lisbon, March, 1998

3. Carvalho, P., Trancoso, I.M., and Oliveira, L.C., "Automatic Segment Alignment for Concatenative Speech Synthesis in Portuguese", *Proc. of the 10th Portuguese Conference on Pattern Recognition, RECPAD'98*, pages 221-226, Lisbon, March, 1998

4. Moulines, E. and Charpantier, F., "Pitch-synchronous waveform techniques for text-to-speech synthesis using diphones", *Speech Comunnication 9*, 453-467, 1990

5. Campbell, N. and Black, A. "CHATR: a multi-lingual speech re-sequencing synthesis system", in *Proc. of Institute of Electronic Information and Communication Engineers-89,* Tokyo, Japan

6. Ribeiro, C., Trancoso, I.M., and Viana, M.C. - "EUROM.1 Portuguese Database", *Report D6 ESPRIT Project 6819 SAM_A (Speech Technology Assessment in Multilingual Applications)*, 1993.

7. Isabel Trancoso, M.Céu Viana, Luís C. Oliveira, M. Isabel Mascarenhas, Pedro Carvalho, Carlos Ribeiro - "Relatório de Execução Material - BDFALA - Base de Dados Falada para o Português Europeu, Projecto PLUS/C/LIN/801/93", JNICT, June 1997 (in Portuguese).

8. Conkie, A. and Isard, S., "Optimal Coupling of Diphones", *2nd ESCA/IEEE Workshop On Speech Synthesis*, Sept. 1994

9. Takeda, Kazuya, Abe, Katsuo and Sagisaka, Yoshinori, "On the basic scheme and algorithms in non-uniform speech synthesis", *Taking Machines: Theories, Models and Designs*, G.Bailly, C.Benoît, and T.R.Sawallis editors, Elsevier Science Publishers B.V., pag.93-105, 1992

10. Shikano, Kiyohiro, and Itakura, Fumitada, "Spectrum Distance Measures for Speech Recognition", *Advances in Speech Signal Processing*, Marcel Dekker, Inc, Chapter 14, pag. 419-452

11. Campbell, W.N., “CHATR: A High-Definition Speech Re-Sequencing System”, *Proc. 3rd ASA/ASJ Joint Meeting*, pages 1223-1228, Hawaii, 1996

12. Moulines, E. and Charpentier, F., “Pitch-Synchronous Waveform Processing Techniques For Text-To-Speech Synthesis Using Diphones”, *Speech Comunication 9*, pages 453-467, 1990

13. Ljolje, Andrej, Hirschberg, Julia and van Santen, Jan P.H., “Automatic Speech Segmentation for Concatenative Inventory Selection”, *Progress in Speech Synthesis*, Springer-Verlag, pages 304-311, 1997

[i] ACKNOWLEDGEMENTS: The work of Pedro Carvalho was sponsored by the grant PRAXIS XXI/BD/4526/94.