Automatic Segment Alignment for Concatenative Speech Synthesis in Portuguese

Pedro Carvalho, Isabel Trancoso, Luís Oliveira

INESC/IST
Rua. Alves Redol, 9, 1000 Lisboa, PORTUGAL
email: Pedro.Carvalho@inesc.pt, Isabel.Trancoso@inesc.pt, Luis.Oliveira@inesc.pt

Abstract: *Concatenative Text-To-Speech synthesizers join pre-recorded segments of speech data in order to produce high quality output speech. The synthesizer has to find the best segment to concatenate from an inventory of speech material. In order to do that, the inventory should be built from a correctly transcribed and time aligned speech database. This paper describes the construction of an automatically alignment tool using a Hidden Markov Model using very small training and test sets.*

Keywords*: Alignment, Segmentation, Concatenative, Synthesis, Speech, HMM.*

1. INTRODUCTION

A generic text-to-speech system can be divided into two main modules: the first one performs text normalization and linguistic processing, and the second one generates the output speech waveform, using as input the string of phonetic symbols and prosodic parameters produced by the first module.

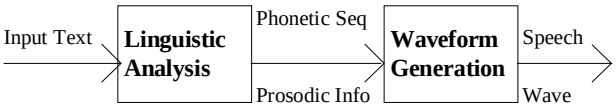


Figure 1. A simple two module decomposition of a generic text-to-speech system

The system that we are currently developing uses the first module of the DIXI [1] system, and substitutes the second module (a rule-based Klatt formant synthesizer) by our own concatenation module.
This module will be responsible for, given an input sequence of phonetic symbols duration and f0 marks, selecting the best matches from an inventory of pre-recorded speech units, and concatenating the selected units to generate the output speech waveform.

The ultimate goal of this work is to construct a concatenative text-to-speech system for European Portuguese that could be trained by any speaker to "talk" with his/her voice. 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.

The first task in building such a system is to create an inventory. Generally, this is achieved by manually transcribing and segmenting a spoken corpus recorded with high quality in a controlled environment, containing phonetically rich sentences and words, in order to contemplate as many different phonetic contexts (left and right) as possible. This task, however, is extremely time-consuming and requires a profound knowledge of phonetics to accurately time align the transcription labels. Hence, automatic segmentation / alignment systems are usually adopted to speed up this procedure.
The automatic system we are planning to develop has two stages. The first stage uses the phonetic transcription module of the DIXI system in order to generate a string of phonetic symbols for the text corresponding to each recorded sentence. The second stage adopts Hidden Markov Models (HMM) to time-align the speech sentences. The development of this alignment tool is the goal of the present work.

This paper is structured as follows: in section 2, we will discuss the training and test sets used to construct the alignment tool. Section 3 will describe the training stage, with an emphasis on the compromises needed in terms of HMM phone model selection for small amounts of training data. After a brief explanation of our software for computing alignment scores, section 4 will show the results obtained in training our tool. Also

discussed are the recognition and alignment results on the test set. Finally, section 5 will comment on the results obtained and discuss the planned future work.

2. TRAINING AND TEST SETS

The spoken corpus used in this work has been recorded in the scope of the European project SAM_A [2] (Speech Technology Assessment in Multilingual Applications). For training the tool, we have selected a subset of this corpus, consisting of 15 passages of 5 sentences each, spoken by one female and one male speaker, amounting to around 4000 PLUs (Phone Like Units) each. The test set consists of one additional passage and 5 filler sentences for the same speakers, amounting to around 950 PLUs each. All this material was transcribed and time-aligned by expert phoneticians, according to the rules specified in [3].

3. THE ALIGNMENT TOOL

The core of the alignment tool is a speaker dependent HMM monophonic network consisting of 60 PLU models. 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.

PLU	Female	Male	PLU	Female	Male
@	26	47	{	304	338
E	11	12	_E	29	75
J	20	32	_O	21	39
L	7	8	_a	137	151
N	111	115	_e	43	57
O	9	21	_I	79	86
R	29	33	_o	53	74
S	128	144	_u	21	29
Z	57	66	_{	28	40
a	31	42	b0	36	42
b	23	28	d0	191	215
d	121	176	e~	25	26
e	0	5	g0	29	38
f	34	44	i~	15	17
g	15	20	j~	21	29
i	96	98	k0	124	131
j	69	89	l~	23	37
k	124	133	o~	13	30
l	66	67	p0	95	119
m	123	142	r0	186	225
n	79	85	t0	198	256
o	9	0	u~	14	14
p	95	119	w~	18	22
r	167	194	{~	20	28
s	130	155	_e~	32	52
t	189	250	_i~	15	15
u	116	140	_o~	17	32
v	72	94	_u~	6	0
w	35	38	_{~	47	61
z	70	71	Sil	123	133

Table 1. PLU count for both sets of each speaker

Table 1 shows the total number of PLUs existing in both sets and for both speakers. We used a modified/extended SAMPA phonetic alphabet. The modifications are: the "_" (underscore) symbol is used as a primary stress mark and the "{" is used in place of "6". We also extended the set to include the stop and burst stages of the voiced and unvoiced stops (p0-p, t0-t, k0-k, d0-d, g0-g and b0-b). Finally, the "N" symbol was added to represent the change of nasal influence from a nasal vowel to a stop consonant.

As it can be seen in Table 1, some of the PLUs have an insufficient number of occurrences to complete the HMM training phase. In the case of vowels, or nasal vowels, we adopted an existing unstressed (or stressed) variant of the PLU, to model both the stressed and unstressed ones. In some cases, although a sufficient number of occurrences exists, the HMM initialization tool rejected a few small duration segments. This can be explained by the fact that we are using a 3-state 3-mixture model with 5ms input vector rate, and small duration segments (<10ms) can not correctly initialize such a complex model. Although not demonstrated in this paper, the use of three Gaussian mixture models improves alignment (and recognition) results, in spite of increasing the number of rejected (small duration) segments.

After the initialization and re-estimation stages, the training proceeded by applying embedded re-estimation followed by Viterbi alignment. Special software was developed to compute the alignment results, comparing the output label files from the Viterbi aligner with the manually aligned label files. In parallel, we also performed Viterbi recognition in order to try to correlate the alignment and recognition results.

In the embedded re-estimation phase, we used a special stop criterion for the number of iterations to perform. The criterion simply consists of stopping this phase when the maximum absolute difference between the time-aligned labels produced in two consecutive iterations is less or equal to the input vector rate (5ms).

The combined re-estimation and embedded re-estimation phases are performed in two different stages. In the first stage, the re-estimation is performed in a single pass and the embedded re-estimation is iterated until the stop criterion is met. The resulting alignment tool is therefore tuned to produce the best results for the training set. In the second stage, the re-estimation and embedded re-estimation are both iterated, but re-estimation uses as input the output labels of the previous iteration. The same stop criterion is applied to the process. The resulting alignment tool is therefore theoretically more "stable". This two-stage process, illustrated in figure 2, attempts remove the dependency of the tool on the small size training set.

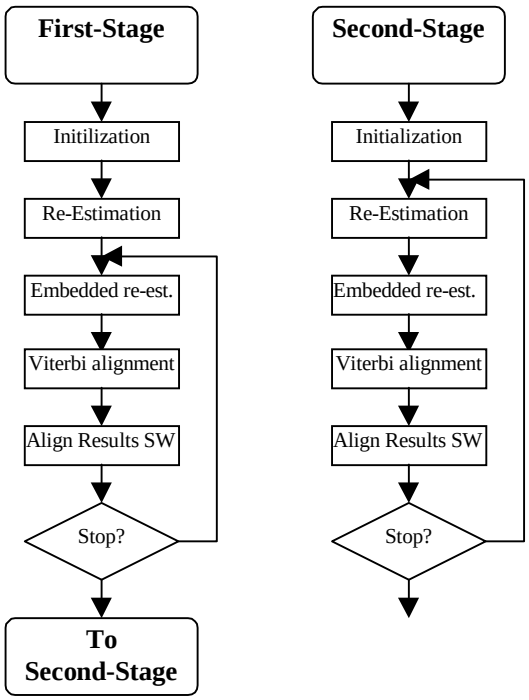


Figure 2. Two-stage process for re-estimation and embedded re-estimation training

4. RESULTS

As stated earlier, a special program was developed to compute the performance of the alignment tool. The program computes the temporal differences between two sets of time-aligned label files, and produces several statistics. For each PLU transition, the maximum positive, maximum negative, average, absolute and RMS difference between the time-aligned labels are computed. Several histograms are also computed allowing the study of four other performance scores, like the percentage of cases with average error less than 10, 20, and 30ms. Some of the scores frequently used by other authors [4][5][6][7] are: the average difference, the RMS difference and the difference for 90% of the cases. Our program also groups the transitions in classes using the following table:

sil	silence	sil
V	Vowels	@ { _{ u _u e _e i _i o _o O _O E _E a _a
NV	Nasal Vowels	{~ _{~ i~ _i~ e~ _e~ u~ _u~ o~ _o~
G	Glides	j w w~ j~
VS	Voiced Stops	d0 d g0 g b0 b
US	Unvoiced Stops	t0 t k0 k p0 p
VF	Voiced Fricatives	v Z z
UF	Unvoiced Fricatives	S f
L	Liquids	r0 R r L l l~
N	Nasals	m n J N

Table 2. PLU classes

One special feature of the alignment results software is to recalculate all the above mentioned scores after adjusting the difference between each PLU transition by subtracting the average transition difference. Also the average difference transition matrix is dumped to a file. We will use this feature latter in this paper.

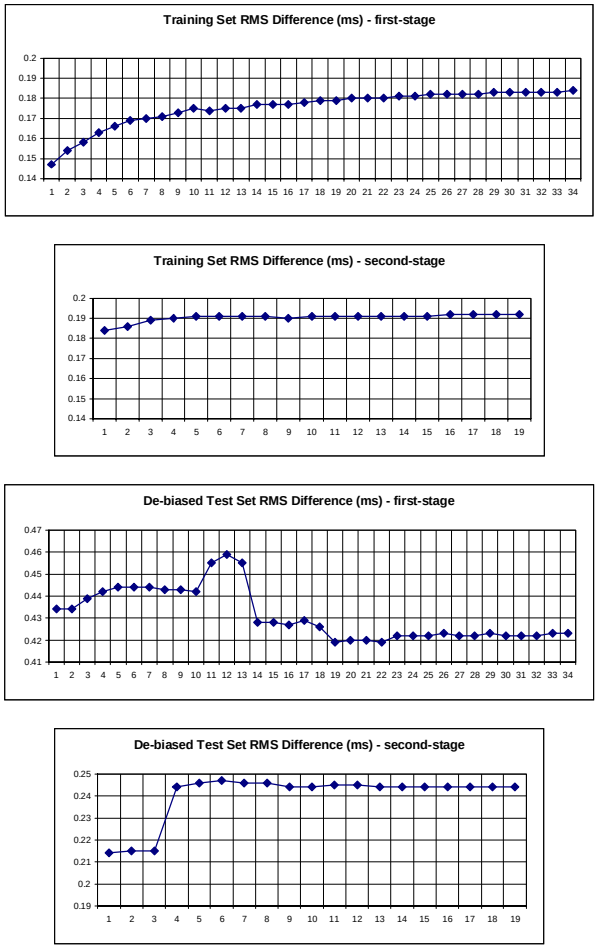


Figure 3. Alignment results for the female speaker.

As it can be seen in Figures 3 and 4, the training phase of the alignment tool for the female and male speakers respectively, converges for higher differences. In other words, the automatically aligned labels converged to values progressively more "distant" from the original (manually aligned) ones. One possible explanation is that the training stages perceive different features from the cepstral input vector than those used by the expert phoneticians that generally adopt time-domain, LPC spectral analysis and formant based criteria.

Hence, if we could compensate for this "distance", the resulting alignment tool would, in principle, be more accurate, i.e. would better approximate the manual alignment accuracy.

We used the above mentioned average difference transition matrix (produced with the alignment results software on the training set) to correct the Viterbi aligned labels, as the last stage of the alignment tool. The subtraction of the average transition difference may be viewed as a de-biasing process. Although used by other authors [5], this process is applied in our system in a context dependent environment, since the average

difference transition matrix allows a right-context dependency lookup.

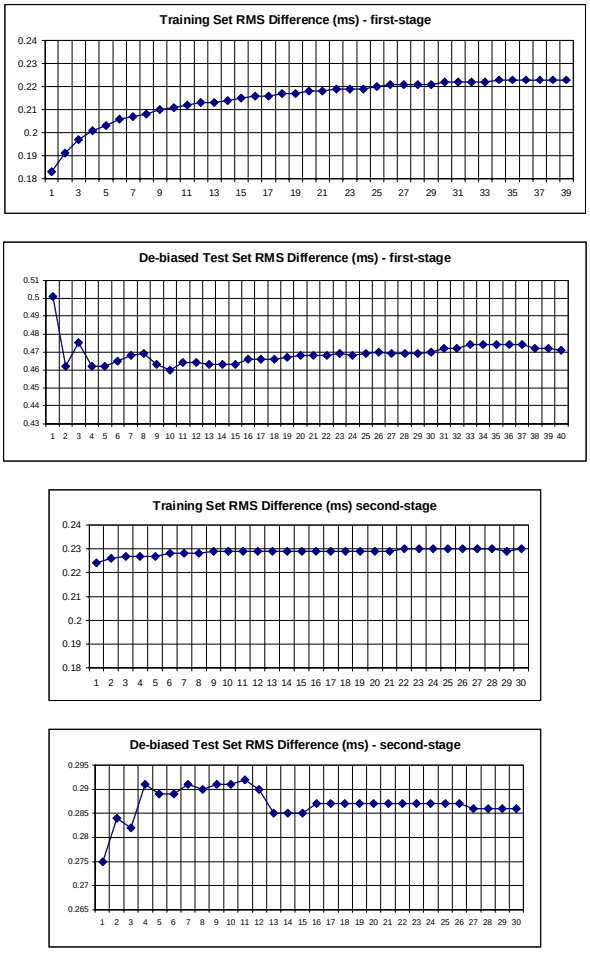


Figure 4. Alignment results for the male speaker

Several other heuristics were tried, although without significant success, in order to improve the performance of the aligner. For instance, a strategy based in pitch-synchronous alignment correction and/or zero-crossing correction was tried but the improvements were not significant and these processes are generally computational intensive.

Figures 5 and 6 illustrate the evolution of the correctness and accuracy scores for PLU recognition. Although a recognition tool is not the present goal of our work, it is interesting to compare figures 5 and 6 with figures 3 and 4, respectively.

The figures show that, in the first stage, the recognition results improve with the number of iterations, and the alignment tool scores stabilize to an offset value, as discussed.

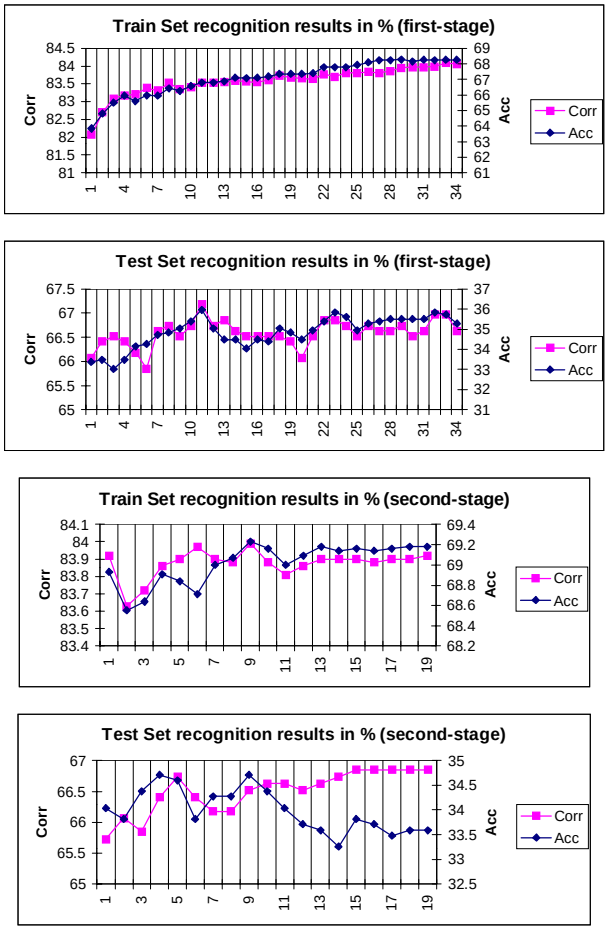
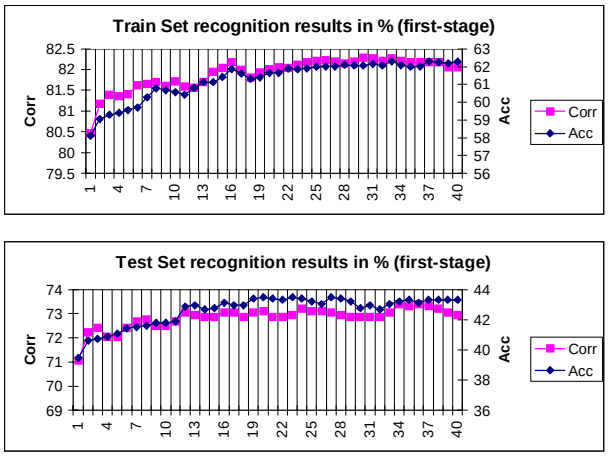


Figure 5. Recognition results for the female speaker

In the second stage, some performance degradation can be observed with the growing number of iterations, namely in terms of accuracy scores. This is consistent with the philosophy behind the second-stage training, since it “flattens” the decision areas of the HMM models, in order to try to stabilize the alignment results.



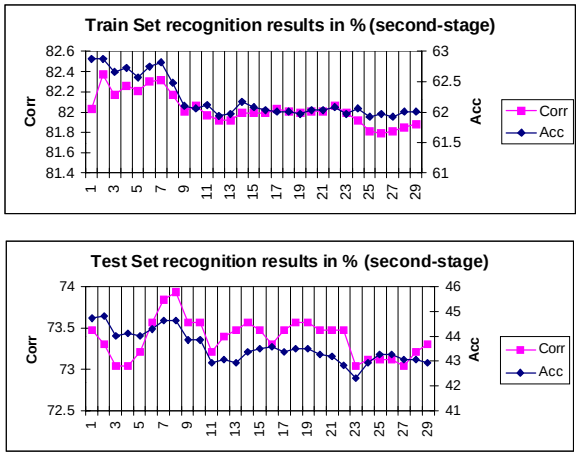


Figure 6. Recognition results for the male speaker

Tables 3 to 8 illustrate the transition class matrix differences on the test set for the male and female speaker before and after applying the de-bias process explained above. These tables should be read from column to row.

	sil	V	NV	G	VS	US	VF	UF	L	N
sil		12		1		2	1	5		1
V	10	9	2	14	32	55	27	25	29	38
NV	2	2	1	1	2	8	3	13	5	10
G	3	12	7		4	2		1	4	2
VS		30	2	5	39	2	4	2	1	9
US		39	1	1	3	92	3	17	4	17
VF	2	22	4	3	4	1	1		4	1
UF		40	4	4	6	4	1	3	5	2
L	1	39	1	2	2	9	0	1	22	0
N	4	36	25	4	2	2	2	2	3	2

Table 3. Total number of transitions on the test set for the female speaker (read from column to row: Vowel followed by silence = 12 occurrences).

ms	sil	V	NV	G	VS	US	VF	UF	L	N
sil		12		8		22	18	18		28
V	26	76	16	54	16	6	10	10	30	18
NV	24	16	4	4	12	6	14	6	26	20
G	8	18	60		4	18		2	34	10
VS		12	6	6	14	98	8	24	0	46
US		12	6	10	22	14	20	10	36	14
VF	14	24	34	22	12	0	10		26	4
UF		14	14	28	22	16	78	90	14	14
L	8	36	40	22	16	12		6	18	
N	18	18	14	14	6	12	10	6	24	22

Table 4. Maximum error for 90% of the cases of the test set classes for the female speaker

ms	sil	V	NV	G	VS	US	VF	UF	L	N
sil		14		24		12	2	20		8
V	16	66	16	58	16	8	10	6	28	14
NV	24	16	4	14	12	6	12	6	12	18
G	8	22	60		4	8		2	34	10
VS		10	28	8	12	98	12	24	2	34
US		12	6	10	18	10	20	6	38	20
VF	10	10	34	22	12	0	2		36	4
UF		8	14	22	30	14	78	90	18	2
L	0	36	54	22	16	26		8	18	
N	6	22	20	8	14	12	4	2	32	22

Table 5. Maximum error for 90% of the cases of the test set classes for the female speaker after de-biasing

	sil	V	NV	G	VS	US	VF	UF	L	N
sil	0	11	0	1	1	2	0	9	1	1
V	9	11	1	9	54	51	21	31	49	38
NV	0	2	0	2	4	8	5	15	11	11
G	0	18	9	0	0	4	1	1	2	0
VS	2	49	9	2	71	3	6	6	11	6
US	3	41	5	7	4	99	0	12	9	20
VF	1	28	0	5	6	1	0	0	2	0
UF	7	33	5	5	13	6	2	4	5	2
L	2	52	2	2	5	20	2	3	57	3
N	2	29	27	2	7	6	6	1	1	1

Table 6. Total number of transitions on the test set for the male speaker.

ms	sil	V	NV	G	VS	US	VF	UF	L	N
sil		98		44	98	70		20	98	38
V	26	32	38	98	20	10	30	10	28	16
NV		36		4	12	8	16	12	12	16
G		30	46			48	14	6	48	
VS	24	30	32	22	18	2	46	32	26	66
US	44	16	10	18	10	14		32	42	20
VF	6	24		26	24	10			22	
UF	16	14	22	28	8	10	36	56	14	6
L	18	42	38	30	20	14	22	10	14	32
N	18	16	26	12	22	24	30	20	6	0

Table 7. Maximum error for 90% of the cases of the test set classes for the male speaker

ms	sil	V	NV	G	VS	US	VF	UF	L	N
sil		54		44	74	26		10	98	20
V	16	14	8	98	18	8	24	6	20	12
NV		4		24	0	12	10	6	16	14
G		34	38			42	14	0	48	
VS	10	18	16	20	14	4	48	34	26	64
US	44	8	10	6	10	12	32	38	18	
VF	12	26	10	24	6				2	
UF	10	10	20	16	6	10	36	56	18	2
L	8	30	38	30	10	8	14	6	10	34
N	14	16	22	12	12	12	32	6	0	6

Table 8. Maximum error for 90% of the cases of the test set classes for the male speaker after de-biasing

As it can be seen in tables 3 and 6, the test set does not cover all the transition classes evenly. In some cases, there is an insufficient number of occurrences to try to conclude any special or abnormal behavior of the alignment tool.

Nevertheless, comparing tables 4 and 5 for the female speaker and 8 and 7 for the male speaker, the merit of de-biasing is obvious.

If we ignore the worst results produced by the alignment tool in cases with a very low number of occurrences, the most notorious errors occur in the transitions: vowel-vowel, nasal vowel-glide, glide-vowel and nasal vowel-liquids, for the female speaker.

In the case of the male speaker, an abnormal error is reported in almost all transitions to silence. For this fact certainly contributes the tendency of the male speaker to produce very low energy PLUs at the end of a sentence. Removing these abnormal errors, the worst case transitions for the male speaker: glide-vowel and nasal-voiced stop.

Some of the errors obtained with stop consonants, for both speakers, are due to their small duration in most of the cases.

Normal / De-biased	RMS	<10ms	<20ms	<30ms	90%
Female	0.51 ms 0.45 ms	66 % 74 %	89% 89%	95% 96%	21 ms 21 ms
Male	0.54 ms 0.44 ms	62% 73%	84% 90%	93% 95%	25ms 20ms

Table 9. Alignment tool overall performance scores for each speaker, before and after applying the de-biasing process

5. CONCLUSIONS AND FUTURE WORK

The study of the correlation between the alignment scores and the recognizer scores is inconclusive. Again the small amount of training data possible does not contribute to clarify the problem. Our best guess is that there is no correlation between a good recognizer and a good aligner.

The baseline HMM alignment tool yielded in 90% of the segments an error below 25 ms, for both speakers in the test set. These values are comparable with other published work containing more extensive training sets, and in some case more complex HMM networks. The de-biasing process enabled us to obtain further improvements in some of the scores

Our alignment tool produces worst results in vowel-vowel, glide-vowels and nasal vowel-glide transitions. These transitions are the ones that expert phoneticians are more likely to disagree, like explained in [5].

The bad results obtained in glides, suggest that several diphthong HMM models should be tried instead of decomposing the diphthong in their vowel and glide parts.

Nevertheless, one question arises: does a concatenation synthesizer need more accuracy in the time-align speech-database it uses to construct its inventory? The answer could be no, specially if the inventory build tool uses some algorithm to try to find best concatenation boundary, for instance a spectral mismatch based algorithm.

The transcribed subset of the SAM corpus we have used for training and testing the alignment tool is much too small to construct a proper inventory of acoustic segments. Our future efforts will be directed towards augmenting this corpus using the spoken corpus recently collected in the scope of the national project BDFALA [3] (sponsored by the Lusitânia project, JNICT). This will be achieved by using the current alignment tool in a semi-automatic segmentation of this larger corpus. The time-aligned corpus can then be used to retrain the present aligner, in a process generally known as bootstrap.

6. REFERENCES

[1] L. C. Oliveira, M. C. Viana, and I. M. Trancoso - «A rule-based text-to-speech system for portuguese»,. In Proc. Int. Conf. on Acoustic Speech and Signal Proc., volume 2, pages 73--76, São Francisco, March 1992.

[2] C.Ribeiro, I.Trancoso, C.Viana - «EUROM.1 Portuguese Database», Report D6 ESPRIT Project 6819 SAM_A (Speech Technology Assessment in Multilingual Applications), 1993.

[3] Isabel Trancoso, M.Céu Viana, Luis 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.

[4] Colin W. Wightman, David T. Talkin, The aligner: Texto-to-Speceh Alignment Using Markov Models, Progress In Speech Synthesis, pages 313-323

[5] Andrej Ljolje, Julia Hirschberg, Jan P.H. van Santen, Automatic Speech Segmentation for Concatenative Inventory Selection, Progress In Speech Synthesis, pages 304-311

[6] Andrej Ljolje, Micheal D. Riley, Automatic Segmentation Of Speech for TTS, EUROSPEECH'93 pages 1445-1448, Volume 2, September 1993

[7] O. Boeffard, L. Miclet, S. White, Automatic Generation of Optimized Unit Dictionaries for Text To Speech Synthesis, Proceedings ICSLP 92, pages 1211-1214, Volume 2, October 1992

Acknowledgements

The authors would like to acknowledge the collaboration of Dr.M.Céu Viana and M.Isabel Mascarenhas from CLUL, for their indispensable contribution to this work.

The present work is part of Pedro Carvalho's PhD thesis, "Automatic Segment Determination for Portuguese Concatenative Speech Synthesis", sponsored by a JNICT scholarship (PRAXIS XXI / BD / 4526 / 1994).