AUTOMATIC PHONEME SEGMENTATION WITH RELAXED TEXTUAL CONSTRAINTS

PIERRE LANCHANTIN, ANDREW C. MORRIS, XAVIER RODET, CHRISTOPHE VEAUX

Very high quality text-to-speech synthesis can be achieved by unit selection in a large recorded speech corpus [1]. This technique uses some optimal choice of speech units (e.g. phones) in the corpus and concatenates them to produce speech output. For various reasons, synthesis sometimes has to be done from existing recordings (rushes) and possibly without a text transcription. But, when possible, the text of the corpus and the speaker are carefully chosen for best phonetic and contextual covering, for good voice quality and pronunciation, and the speaker is recorded in excellent conditions. Good phonetic coverage requires at least 5 hours of speech. Accurate segmentation of the phonetic units in such a large recording is a crucial step for speech synthesis quality. While this can be automated to some extent, it will generally require costly manual correction. This paper presents the development of such an HMM-based phoneme segmentation system designed for corpus construction.

1. Architecture of the System

The segmentation system presented here is based on the Hidden Markov Models Toolkit (HTK [2]). It has been designed to perform a Viterbi decoding based on a phoneme-level graph which topology depends of the text transcription availability :

- when a text transcription is not available, a *phoneme bigram language model* is used. Indeed, the absence of script-derived constraints on the realisable phoneme sequences should allow better phoneme recognition for this case. However, the segmentation is less robust to non-speech noises like lipsmack or breathing which can be intermingled with language phonemes;
- when a text transcription is available, the textual information, which can be seen as a constraint on the permissible phoneme sequences, is provided by a *multi-pronunciation phonetic graph*. This graph is built by using a version of Lia_phon, a rule based French text-to-phone phonetisation program [3], which we have improved for this purpose. The graph is built as presented on Fig 1.

Given the graph which has been selected (depending of the text transcription availability), its associated set of HMMs and an acoustic observation (MFCC), the Viterbi decoder [4] then finds the most likely sequence of phonemes given the acoustic signal. Finally, the recognizer outputs the phonetic sequence that best matches the speech data.

2. HMMs Design and Training

2.1. **Text and recording.** The recorded text is a set of 3994 sentences in French, chosen in [5] for good phonetic and contextual covering. A subset of 354 of these sentences has been hand segmented, and then divided into one set of 200 sentences for the tuning of the models (*development set*), and another set of 154 sentences for testing (*test set*). The remaining 3640 sentences are used for model training (*training set*). The acoustic features used in all experiments are Mel-Frequency Cepstral Coefficients (MFCC), together with their first and second smoothed time difference features (which we name MFCC- Energy Delta Acceleration (MFCC-EDA)), calculated on 25 ms sample windows every 5ms.

2.2. **Training procedure.** HMMs used for each phoneme have the same topology of forward and self connections only and no skips. The begin/end silence has a skip from the first to the last active state and vice versa. Since the realisation of phonemes is highly context dependent, the choice of triphones as phonetic units is expected to give better match accuracy than monophones, but not necessarily a better timing accuracy. Then, the two types of model have been experimentally compared in our system. Different numbers

of states per HMM and Gaussians per state were also tested. At first, monophone HMMs are estimated by embedded training on the *training set* using phonetic transcription of the whole each sentence. The phoneme bigram language model is then trained on the corpus phonetic transcription. If aiming for a final monophones based architecture, then a number of steps of *mixture splitting*, are applied while increasing the number of Gaussians per state by splitting the largest Gaussians, and models are re-estimated. If aiming for a final triphone based architecture, then initial triphone models are first obtained from 1-Gaussian monophone models. A clustering procedure is then used to map triphones absent from the corpus onto models of triphones present in the corpus [2]. Several iterations of mixture splitting and re-estimation steps are then applied, as in the case of monophone models.

2.3. **HMM design for phonetic decoding.** Design of the models were conducted on the *development set* sentences for different numbers of Gaussians. Optimisations have been made considering the phoneme bigram language model for which recognition results are more sensitive in the HMMs topology than when considering the multi-pronunciation phonetic graph. HMMs topology is optimised according to segmentation accuracy which is measured by the *Match Accuracy measure* [6]. Initial tests use a model with 3 states, and 1, 2, 3, or 5 Gaussians. Several variations have been tested in these initial tests and the best system configuration for the database was the following :

- 64Hz low-frequency cutoff;
- EDA with 13 base MFCCs;
- Shift of the 25ms sample window;
- Initial training using hand-segmented data.

From this configuration, we then varied both the number of Gaussians per state , the number of states per HMM and the number of Baum-Welch iterations per processing step (3, 6, 9). Figures 2a and 2b show match accuracy (ignoring timing information) according to, respectively, the number of states per model and the number of Baum-Welch iterations per training step. A number of points can be drawn from these Figures:

(1) Figure 2a shows that triphone models generally outperform monophone models for a given number of Gaussians per model. However, overall model size is usually much larger for triphones than for monophones. Further monophone tests would be required to check whether monophone performance will peak at a value lower than peak triphone performance (93.2%). However, as triphone models take account of known context dependencies, it would be expected that triphone model accuracy would have a greater potential to increase as the proportion of triphones represented in the training data increases;
(2) Performance peaks at 7 states per HMM for both monophone and triphone models (9-state performance is almost the same, but slightly worse);
(3) Figure 2b shows that triphone performance peaks at 5 Gaussians per state. Monophones peak somewhere above 40 Gaussians per state.

3. Segmentation using textual knowledge

We then used the topology found in the last section (triphone models, 7 states per phoneme and 5 gaussians per state) and we studied the results obtained with the test set considering the phoneme bigram language model to decide which pronunciations will be allowed in the multi-pronunciation phonetic graph. We then compared the results in term of phoneme recognition precision and phonem boundary precision.

3.1. **Phoneme recognition precision.** Table 1 shows the phoneme confusion counts for the segmentation of the test set based on phoneme bigram language model considering the best model topology presented in the last section (diagonal on the right under Diag). Every one of the errors made was inspected and phonetic rules were deduced from them and incorporated in Lia_phon in order to take the text information into account via multi-pronunciation phonetic graphs. The phoneme confusion matrix resulting from this new segmentation is given in Table 2. Most of the errors are avoided and the match accuracy is now equal to 96.8% compared to 93.2% in the case of the segmentation based on the phoneme bigram language model.

3.2. **Phoneme boundary precision.** We also compared the results in term of *Timing accuracy*. Table 3 shows boundary precision in terms of Timing Accuracy for the segmentation based on the phoneme bigram language model. It also shows the percentage of sentences with all boundaries within tolerance. Two estimated boundaries are not allowed to be matched to the same given boundary. The residual 5% inaccuracy for a tolerance of 70 ms is therefore mostly due not to inaccurate boundary positions, but to extra inserted boundaries (which may in some cases not really be errors, because the hand labelling is not 100% correct). On looking at Table 4, we can see that there is a slight improvement concerning the segmentation precision.

4. Conclusion

This paper has presented some tests and improvements of an HMM-based phoneme segmentation system aimed at the construction of large speech synthesis corpus. Optimal HMM architecture and parameter values have been determined for a high quality monospeaker recording. Segmentation based on phoneme bigram language model, i.e. without text knowledge, and segmentation based on multi-pronunciation phonetic graph with text knowledge, have been studied and allow Match accuracy rates up to, respectively, 93,2% correct phoneme recognition and 96,8% correct labelling. These results suggest that the cost of manual verification and correction of the corpus can be largely reduced. This system is actually tested on a bigger database called "Chronic" which consist of 3 hours of manually segmented speech.

References

[1] R.E. Donovan, "Current status of the IBM trainable speech synthesis system," in *Proc. ESCA Workshop on Speech Synthesis*, Scotland, UK, 2001.

[2] S. Young, G. Evermann, D. Kershaw, G. Moore, J.Odell, D. Ollason, V.Valtchev, and P.Woodland, *The HTK Book*, Cambridge University, 2002.

[3] F. Béchet, "Lia_phon : un système complet de phonetisation de textes," *Traitement Automatique des Langues*, vol. 42, no. 1, pp. 47–68, 2001.

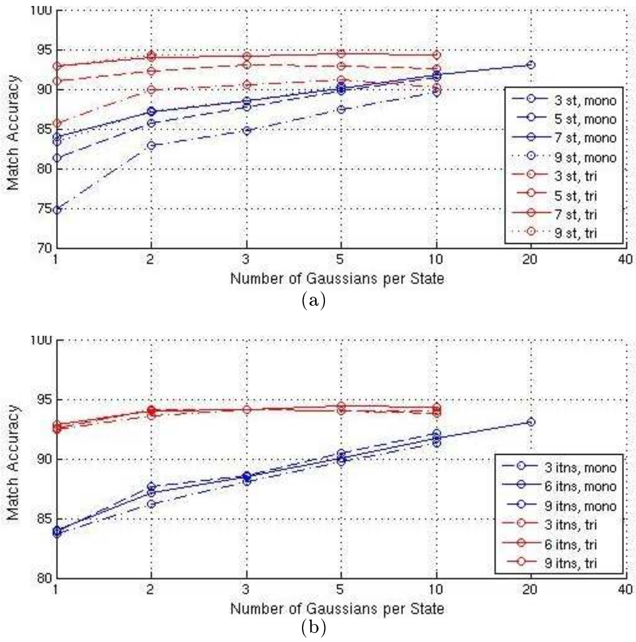
[4] G. D. Fornay, "The Viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–277, 1973.

[5] "Corpatext 1.02," *www.lexique.org/public.Corpatext.php*.

[6] A.C. Morris, V. Maier, and P. D. Green, "from WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition," in *Proc. ICSLP*, 2004.

```
1)          Le              petit            chat
2)       l[@|2|9]        p[@|2|9]ti           Sa
3)      {l[@|2|9]}      {p[@|2|9]ti}         {Sa}
4)  sil {l[@|2|9]}[sp]{p[@|2|9]ti}[sp]{Sa} sil
```

Figure 1. *Phonetic graph construction for the sentence "le petit chat". The metacharacters : | denotes alternatives, [ ] enloses options, { } denotes zero or more repetitions: 1) the sentence is splitted in words, 2) phonetisation of each word, 3) each word is optional and can be repeated, 4) optionals sp are added between words.*



(a)

(b)

Figure 2. *a) Match accuracy versus number of Gaussians per state for various number of states per HMM (6 iterations per step, mono or triphones). b) Match accuracy versus number of Gaussians per state for various number of training iterations per step (7 state HMMs, mono or triphones)*

TABLE 1. *Phoneme confusion matrix obtained after the test set segmentation based on the phoneme bigram language model (row=true phoneme, column=phoneme identified, diagonal on the right under Diag). Hits=3000 Deletions=31 Substitutions=103, Insertions=85, Accuracy=93.01%, Match accuracy=93.20%.*

	s	s	a	m	E	e	b	R	9	n	y	k	u	s	t	o	i	z	l	2	f	a	d	w	0	v	p	@	9	e	j	g	o	H	Z	S	Del	Ins	Diag
	i	p								~											~										~		~							
sp	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	7	40
a	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	9	170
E	.	.	.	.	12	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	3	93
e	.	.	.	.	8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	3	115
R	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	1	192
9~	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	1	1	.	.	.	.	.	.	.	.	.	.	.	1	1	67
y	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	2	.	.	.	.	.	.	1	.	.	.	.	.	2	.	.	.	.	.	.	.	.	1	0	72
s	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	2	113
t	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	3	146
o	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	1	.	.	.	.	3	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	41
i	.	.	.	.	.	1	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	2	141
z	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	0	50
l	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	4	191
2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	8	5	.	.	.	.	.	.	.	.	.	.	.	.	.	0	1	54
w	.	.	.	.	.	.	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	0	23
0	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.	.	.	.	.	.	.	.	4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	1	54
p	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	9	95
@	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	29	116
9	.	.	1	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	0	18
e~	.	.	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	0	44
j	.	.	.	.	.	.	.	.	.	.	.	.	.	.	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	2	45
o~	.	.	.	.	.	.	.	.	1	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	0	68

TABLE 2. *Phoneme confusion matrix obtained after the test set segmentation based on the multi-pronunciation phonetic graph (row=true phoneme, column=phoneme identified, diagonal on the right under Diag). Hits=3084 Deletions=16 Substitutions=34, Insertions=50, Accuracy=96.81% , Match accuracy=96.85%.*

	s	s	a	m	E	e	b	R	9	n	y	k	u	s	t	o	i	z	l	2	f	a	d	w	0	v	p	@	9	e	j	g	o	H	Z	S	Del	Ins	Diag
	i	p								~											~										~		~							
sp	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	17	42
a	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	2	170
E	.	.	.	.	.	5	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	0	101
e	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	2	117
y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	0	0	76
k	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	0	79
s	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	2	113
o	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0	0	45
z	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	2	53
2	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	7	.	.	.	.	.	.	.	.	.	.	.	.	.	0	0	59
@	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	10	19	111
j	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	0	49

TABLE 3. *Phoneme boundary precision (left) and whole sentence alignment accuracy (right) for the segmentation based on the phoneme bigram language model with T=number of fully correct, F=number of not fully correct)*

Boundary accuracy						Whole sentence accuracy			
Tol	**TAcc**	**H**	**D**	**I**	**N**	**Acc**	**T**	**F**	**N**
5	19.76	991	1989	2036	2980	0.00	0	154	154
10	43.91	1833	1147	1194	2980	0.00	0	154	154
20	75.54	2585	395	442	2980	6.49	10	154	154
30	87.37	2801	179	226	2980	22.08	34	120	154
50	93.46	2902	78	125	2980	37.66	58	96	154
70	95.16	2929	51	98	2980	48.05	74	80	154
100	96.31	2947	33	80	2980	51.95	80	74	154
500	96.95	2957	23	70	2980	54.55	84	70	154

TABLE 4. *Phoneme boundary detection precision (left) and whole phrase alignment accuracy (right) for the segmentation based on the multi-pronunciation phonetic graph (T=num fully correct, F=num not fully correct) for the segmentation based on the word-pair grammar*

Boudary accuracy						Whole phrase accuracy			
Tol	**TAcc**	**H**	**D**	**I**	**N**	**Acc**	**T**	**F**	**N**
5	21.21	1049	1931	1965	2980	0.00	0	154	154
10	45.80	1883	1097	1131	2980	0.00	0	154	154
20	78.07	2628	352	386	2980	5.19	8	146	154
30	88.79	2819	161	195	2980	22.08	34	120	154
50	94.61	2914	66	100	2980	42.86	66	88	154
70	96.07	2937	43	77	2980	51.95	80	74	154
100	97.17	2954	26	60	2980	58.44	90	64	154
500	97.76	2963	17	51	2980	62.34	96	58	154