Automatic Phoneme Segmentation with Relaxed Textual Constraints

Automatic Phoneme Segmentation with Relaxed Textual Constraints

Pierre Lanchantin

University of Cambridge

38 PUBLICATIONS 256 CITATIONS

SEE PROFILE

Christophe Veaux

The University of Edinburgh

33 PUBLICATIONS 102 CITATIONS

SEE PROFILE

Automatic Phoneme Segmentation with Relaxed Textual Constraints

Pierre Lanchantin, Andrew C. Morris, Xavier Rodet, Christophe Veaux

IRCAM, Analysis-Synthesis Group
1, place Igor Stravinsky, F-75004 Paris, France
{lanchant,rod,veaux}@ircam.fr

Abstract

Speech synthesis by unit selection requires the segmentation of a large single speaker high quality recording. Automatic speech recognition techniques, e.g. Hidden Markov Models (HMM), can be optimised for maximum segmentation accuracy. This paper presents the results of tuning such a phoneme segmentation system. Firstly, using no text transcription, the design of an HMM phoneme recogniser is optimised subject to a phoneme bigram language model. Optimal performance is obtained with triphone models, 7 states per phoneme and 5 Gaussians per state, reaching 94.4% phoneme recognition accuracy with 95.2% of phoneme boundaries within 70 ms of hand labelled boundaries. Secondly, using the textual information modeled by a multi-pronunciation phonetic graph built according to errors found in the first step, the reported phoneme recognition accuracy increases to 96.8% with 96.1% of phoneme boundaries within 70 ms of hand labelled boundaries. Finally, the results from these two segmentation methods based on different phonetic graphs, the evaluation set, the hand labelling and the test procedures are discussed and possible improvements are proposed.

1. Introduction

Very high quality text-to-speech synthesis can be achieved by unit selection in a large recorded speech corpus (Donovan, 2001). 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. We examine in particular two modes of decoding. In the first mode, the decoding is based on a *phoneme bigram language model* without any text knowledge. In the second mode, the decoding is based on a *multi-pronunciation phonetic graph* built according to the text.

We first present the speech database recording, the architecture of the system and the training procedure. We then detail the tests conducted to design the best models considering the segmentation based on the phoneme bigram language model. Finally we discuss the procedure, the results and future work.

2. Text and recording

The recorded text is a set of 3994 sentences in French, chosen in (Corpatext, 2006) for good phonetic and contextual covering. It was read by a male French speaker in a anechoic room and recorded with a high quality microphone and a 16 bits 44.1 kHz analog to digital converter. Instructions for the speaker were to articulate clearly while keeping a natural elocution, not too fast and without too much expressive variation. Each sentence is recorded in a separate file. By pressing a button, the speaker could rerecord any sentence until he was satisfied and then go on to the next. 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.

3. Architecture of System

The segmentation system presented here is based on the Hidden Markov Models Toolkit (HTK (Young et al., 2002)). It has been designed to perform a Viterbi decoding based on a *phoneme bigram language model* when the text transcription is unknown, or to make use of the textual information modeled by a *multi-pronunciation phonetic graph* when the text is at least approximately known. When a text transcription is not available or when the pronunciation of the speaker differs from any of the permissible ones allowed in the multi-pronunciation phonetic graph, 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. More, some phoneme, such as /e/ (s**e**s) and /E/ (s**e**ize), can be intermingled if the models are not accurate enough. Therefore, this scheme requires very accurate phoneme models (with large number of Gaussians per state), which implies sufficient amount of training data for every phoneme and, in the

This research was partially funded by the French RIAM network project VIVOS. The second author is now working at Spinvox Ltd, Marlow, UK. Thanks to G. Gravier for useful discussions and advices.

```
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 french sentence "le petit chat". The metacharacters : | denotes alternatives, [] denotes encloses 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.

case of the choice of triphones as phonetic units, in every left-right phonetic context.

On the other hand, 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 (Béchet, 2001), which we have improved for this purpose. The graph is built as the following and an example of the graph construction is presented on Figure 1:

1. The sentence is splitted into words.

2. For each word, Lia_phon generates different pronunciations from which a corresponding multi-pronunciation phonetic graph is generated.

3. Connections between the begining and the end of a word and vice versa are added to allow skipping or repetition of a word.

4. Optional short pauses are added between words.

Given the graph which has been selected (depending of the text transcription availability), its associated set of HMMs and an acoustic observation (MFCC), the log probability of any path through the graph can be computed. The Viterbi decoder (Fornay, 1973) then finds the path through the graph which maximises the log probability. Finally, the recogniser outputs the phonetic sequence that best matches the speech data.

Two types of model have been experimentally compared in our system. In the case of monophones, a separate HMM is used to model each phoneme. In the case of triphones, a separate HMM is used to model each phoneme in each left-right context. Acoustic variations within each HMM state due to differences in mode of speech, speaker mood, background noises, etc., are modelled by a separate multivariate Gaussian distribution within the Gaussian mixture model (GMM) used to represent the distribution of feature vectors.

4. Training procedure

HMMs used for each phoneme (except "sil" and "sp") 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. Different numbers of states per HMM and Gaussians per state were tested. The models are estimated by embedded training, using a single standard phonetic transcription of the whole sentence, though the transcription text does not fit perfectly to the speech recording, and an improvement will be proposed at the end of the paper. Monophone HMMs are estimated by embedded training on the *training set*. The phoneme bigram language model is then trained on the *training set* phonetic transcription by making the assumption that the corpus is the realisation of a first order Markov chain the states of which are the phonemes of the language.

If aiming for a final monophone based model, 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 model, 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 (Young et al., 2002). Several iterations of mixture splitting and re-estimation steps are then applied, as in the case of monophone models.

5. 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 the *Match Accuracy measure* (Morris et al., 2004) $M_{Acc} = 100 \times H/(H+S+D+I)$ where H, S, D, I are the hit (H), substitution (S), insertion (I) and deletion (D) counts obtained by Viterbi alignment of the given and detected phoneme sequences. This measure ignores timing information completely.

Initial tests use a model with 3 states, and 1, 2, 3, or 5 Gaussians. Variations tested in these initial tests include the following: cepstral mean subtraction of the features; tying of the central state of the inter or intra-word short-silence HMM to that of the beginning and end silence HMM, with given, extra or no forward and backwards skips added to the silence model; low/high-frequency cutoff; different numbers of MFCCs; different sample window sizes and shifts, and initial training of HMMs using hand-segmented data. Finally, 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 (1, 2, 3, 5, 10, 20, 40) , the number of states per HMM (3, 5, 7, 9) and the number of Baum-Welch iterations per processing step (3, 6, 9).

Figures 2 and 3 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. Figure 4 shows match accuracy against model

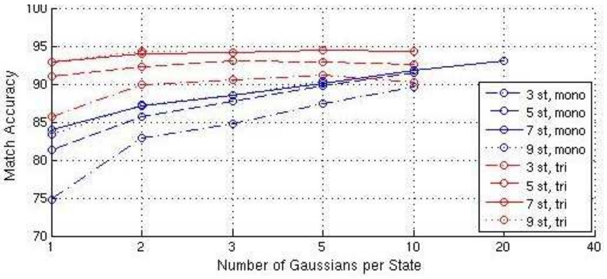

Figure 2: Match accuracy versus number of Gaussians per state for various number of states per HMM (6 iterations per step, mono or triphones)

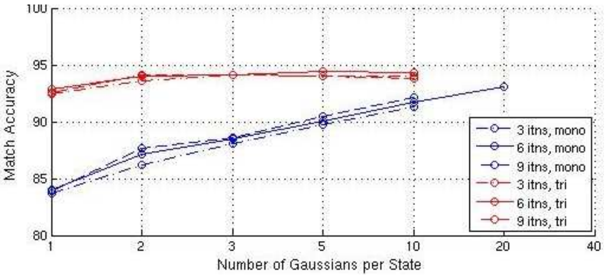

Figure 3: Match accuracy versus number of Gaussians per state for various number of training iterations per step (7 state HMMs, mono or triphones)

size, for monophone and triphone models. A number of points can be drawn from these Figures:

- Figure 2 shows that triphone models generally outperform monophone models for a given number of Gaussians per model in terms of match accuracy. However, overall model size is usually much larger for triphones than for monophones (see Figure 4). Further monophone tests would be required to check whether monophone performance will peak at a value lower than peak triphone performance (94.4%). 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;
- Figure 2 shows that performance peaks at 7 states per HMM for both monophone and triphone models (9-state performance is almost the same, but slightly worse);
- Figure 3 shows that triphone performance start to saturate at 2 Gaussians per state and peaks at 5 Gaussians per state. Monophones peak somewhere above 40 Gaussians per state;
- Performance peaks at 6 Baum-Welch iterations per processing stage. However, performance does not increase very significantly as this number of iterations increases beyond 3.

The results concerning concerning monophones and triphones models with 7 states (6 Baum-Welch iterations per processing stage) considering differents number of Gaussians per state are resumed in the Table 1.

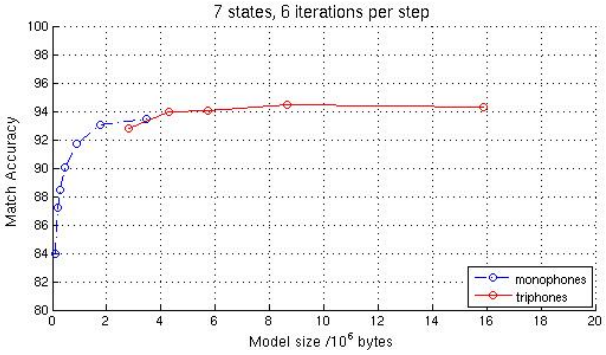

Figure 4: Match accuracy for monophone (1, 2, 3, 5, 10, 20, 40 Gaussian) and triphone (1, 2, 3, 5, 10 Gaussian) models against model size.

NGaussians	1	2	3	5	10	20
Monophones						
MAcc(%)	83.98	87.18	88.44	90.04	91.75	93.06
H	2694	2775	2800	2848	2891	2936
D	145	128	118	105	97	79
S	295	231	216	181	146	119
I	74	49	32	29	18	21
N	3134	3134	3134	3134	3134	3134
bytes/10^6	0.10	0.19	0.28	0.45	0.88	1.75
Triphones						
MAcc(%)	92.83	93.99	94.08	**94.47**	94.30	-
H	2989	3004	3021	3026	3030	-
D	26	26	24	22	24	-
S	119	104	89	86	80	-
I	86	62	77	69	79	-
N	3134	3134	3134	3134	3134	-
bytes/10^6	4.34	6.82	9.22	14.96	28.2	

Table 1: Results concerning monophones and triphones models with 7 states considering different number of Gaussians per state (6 Baum-Welch iterations per processing stage).

6. Segmentation using textual knowledge

We now use the topology found in the last section (triphone models, 7 states per phoneme and 5 gaussians per state) and we study 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 compare the results in term of phoneme recognition precision and phonem boundary detection precision.

6.1. Phoneme recognition precision

Table 2 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 we only show the most relevant ones. Of the remaining errors, listening tests showed that a large proportion were due to in-

```
    s s a m E e b R 9 n y k u s t o i z l 2 f a d w O v p @ 9 e j g o H Z S.
    i p             ~                         ~                 ~     ~
    l                                                                          Del Ins Diag
sil . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   0  308
 sp . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    1   7   42
  a . . . . 1 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   7  169
  m . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   0   81
  E . . . . . 10. . . . . . 1 . . . . . . . . . . . . . . . . . . . . . . . .    2   1   94
  e . . . . 6 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   6  118
  b . . . 1 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   1   68
  R . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    2   0  192
 9~ . . . . . . . . . . . . . . . . . . . . . . . . 1 . . . . 3 . . . . . . .    0   2   69
  n . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   1   79
  y . . . . . . . . . . . . . . . . . . . 1 . . . . . . . . . . . . . . . . .    0   0   78
  k . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   1   79
  u . . . . . . . . . . . . . . . . . . . . . . . . . . 1 . . . . . . . . . .    0   1   77
  s . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   2  114
  t . . . . . . . . . . . . . . . . . . . . . . 3 . . . . . . . . . . . . . .    0   3  146
  o . . . . . . . . . . . . . . . . . . . . . . . . 3 . . . . . . . 1 . . . .    0   3   44
  i . . . . . . . . . . 1 . . . . . . . . . . . . . . . . . . . . . . . . . .    2   1  145
  z . . . . . . . . . . . . . 1 . . . . . . . . . . . . . . . . . . . . . . .    0   2   53
  l . . . . . 1 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    2   2  192
  2 . . . . . . . . . . 1 . . . . . . . . . . . . . . . . 8 . . . . . . . . .    0   4   58
  f . . . . . . . . . . . . . . 1 . . . . . . . . . . . . . . . . . . . . . .    0   0   41
 a~ . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1 . . . .    0   1   67
  d . . . . . . . . . 2 . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   2   79
  w . . . . . . . . . . . . 4 . . . . . . . . . . . . . . . . . . . . . . . .    1   1   21
  O . . . . . . . . . . . . . . . 4 . . . 1 . . . . . . . 4 1 . . . . . . . .    1   2   53
  v . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   1   61
  p . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0  10   96
  @ . . . . . . . . . . . . . . . . . . . . . . . 1 2 . . . 4 . . . . . . . .    5   3  110
  9 . . 2 . . . . . . . . . . . . . . . . 1 . . . . 1 . . . . . . . . . . . .    0   0   18
 e~ . . 3 . . . . . 1 . . . . . . . . . . . . . . . . . . . . . . . . . . . .    2   2   44
  j . . . . . . . . . . . . . . . . 4 . . . . . . . . . . . . . . . . . . . .    2   2   46
  g . . . . . . . . . . . 1 . . . . . . . . . . . . . . . . . . . . . . . . .    0   0   37
 o~ . . . . . . . 1 . . . . . . . . . . . . . 1 . . . . . . . . . . . . . . .    1   0   70
  H . . . . . . . . . . 1 . . . . . . . . . . . . . . . . . . . . . . . . . .    0   0    4
  Z . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    1   0   48
  S . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   1   25
```

Table 2: 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=3026 Deletions=22 Substitutions=86, Insertions=85, Match accuracy=94.47%.

```
    s s a m E e b R 9 n y k u s t o i z l 2 f a d w O v p @ 9 e j g o H Z S
    i p             ~                         ~                 ~     ~
    l                                                                          Del Ins Diag
sil . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   0  308
 sp . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    2  17   42
  a . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   2  170
  m . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   0   81
  E . . . . . 5 . . . . . . . . . . 1 . . . . . . . . . . . . . . . . . . . .    0   0  101
  e . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   2  117
  b . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   0   69
  R . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   0  194
 9~ . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   0   73
  n . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   1   79
  y . . . . . . . . . . . . . . . . 1 . . . . . . . . . . . . . . . . 2 . . .    0   0   76
  k . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   0   79
  u . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   1   78
  s . . . . . . . . . . . 1 . . . . . . . . . . . . . . . . . . . . . . . . .    0   2  113
  t . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   0  149
  o . . . . . . . . . . . . . . . . . . . . . . . . 3 . . . . . . . . . . . .    0   0   45
  i . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    1   0  147
  z . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    1   2   53
  l . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   1  195
  2 . . . . . . . . . . 1 . . . . . . . . . . . . . . . . . 7 . . . . . . . .    0   0   59
  f . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   0   42
 a~ . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   1   68
  d . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   0   81
  w . . . . . . . . . . . . 1 . . . . . . . . . . . . . . . . . . . . . . . .    0   0   25
  O . . . . . . . . . . . . . . . 1 . . . . . . . . . . . . . . . . . . . . .    0   0   63
  v . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   1   61
  p . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   0   96
  @ . . . . . 1 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .   10  19  111
  9 . . . . . . . . . . . . . . . . . . . 1 . . . . . . . . . . . . . . . . .    0   0   50
 e~ . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   0   49
  j . . . . . . . . . . . . . . . . 1 . . . . . . . . . . . . . . . . . . . .    2   0   38
  g . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   1   73
 o~ . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   0   21
  H . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   0    5
  Z . . . . . . . . . . . . . 1 . . . . . . . . . . . . . . . . . . . . . . .    0   0   48
  S . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .    0   0   25
```

Table 3: 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, Match accuracy=96.85%.

Boundary accuracy						Whole sentence accuracy			
Tol(ms)	**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 *phoneme bigram language model* with Tol=Tolerance in ms, TAcc=Timing accuracy in %, H=Hits, D=Deletions, I=Insertions, T=number of fully correct, F=number of not fully correct)

Boudary accuracy						Whole phrase accuracy			
Tol(ms)	**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

Table 5: Phoneme boundary detection precision (left) and whole phrase alignment accuracy (right) for the segmentation based on the *multi-pronunciation phonetic graph* (Tol=Tolerance in ms, TAcc=Timing accuracy in %, H=Hits, D=Deletions, I=Insertions, T=num fully correct, F=num not fully correct)

accurate hand labelling, particularly the *schwa* phoneme /@/. Also, the majority of substitution errors were due to overlapping vowel classes, such as /e/ (s**e**s) → /E/ (s**e**ize) and vice versa, /@/ (null**e**) → /2/ (d**eu**x) and vice versa, /O/ (c**o**mme) → /o/ (gr**o**s) and vice versa, and /w/ (c**o**in) → /u/ (d**ou**x). Another source of errors are repeated phonemes (such as /e e/ in “anne lue”, or /o o/ in “pot au” or “zoo”, where there is sometimes an audible dip in the middle but sometimes just a slightly longer than usual duration. The division between single and double occurrences therefore becomes blurred and this is then reflected by confusion between single and double occurrences in the automatic segmentation. Finally, many insertions are due to *schwas*.

According to these remarks, we incorporated phonetics rules 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 3. Most of the errors are avoided and the match accuracy is now equal to 96.8% compared to 94.4% in the case of the segmentation based on the phoneme bigram language model. Most of the errors are still due to insertion/deletion of the *schwa* phoneme.

6.2. Phoneme boundary detection precision

We also compared the results in term of Timing accuracy which is measured by the *Timing Accuracy measure* (tol) $T_{Acc} = 100 \times H/(H + D + I)$ where H is the number of *given* transitions (i.e. manually placed) matched with a closest *estimated* transition which falls within a given tolerance (tol), D is the number of given transitions not matched with an estimated transition, and I is the number of estimated transitions not matched with a given transition. This measure ignores phoneme identity information (label) completely. Table 4 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 in not 100% correct). On looking at Table 5, we can see that there is a slight improvement concerning the segmentation precision from 95.2% (phoneme bigram language model) to 96.1% (multi-pronunciation phonetic graph) of phoneme boundaries within 70 ms of hand labelled boundaries.

7. Discussion and future work

The match accuracy rate here obtained are promising but should be interpreted with care. When building a corpus from a controlled recording with a text transcription, a match accuracy of 96.8% means that only some 3% of the phoneme labels need hand correction. This will largely diminish the cost of hand manipulations. However, error locations are not provided directly in the system described here. Some confidence measures still needs to be computed such as in (S. Nefti and Moudenc, 2003). Another indication of possible errors could probably come from the comparison of the phoneme sequences provided by the segmentation based on the phoneme bigram language model (no text used) and the segmentation based on the phonetic graph: we will also implement a comparison of these two phoneme sequences as an indication of a possible error. Match accuracy should be taken more as an indication of the influence of the chosen architecture and parameters than ground truth results. This is due to various reasons, among which: the test set is relatively small and we have to increase the hand segmented set. Also, hand segmentation is far from being error proof and needs very careful verification. It is probable that our test set contains errors requiring careful examination.

As mentioned in section 4, the transcription text on the whole sentences does not necessarily fit perfectly with the recorded speech and this can degrade the estimation of the models during the training procedure. To deal with this problem, we propose to gradually relax the textual constraint during training as follows: HMMs parameters are first estimated by embedded training, as described in section 4, using the rule-based phonetisation transcription of the whole sentences given by Lia_phon. After a few steps of estimation, the phoneme models created so far are used to realign the training data according to the multi-pronunciation phonetic graph, and to create new transcriptions that best match the speech recording. Then, the model estimation is refined by embedded training based on the new transcriptions. Finally, the training data are realigned according to the phoneme bigram language model without using any textual constraint. Thus, even if the text does not correspond exactly to the speech recording, one should obtain a good estimation of the phoneme models which finally does not depend on the text.

System test results have been presented for the case of a single speaker. However, the system has also been trained on the Bref-80 multispeaker data base (Lamel et al., 1991) and segmentation has been applied to unknown speaker recordings with promising results. More tests are to be done to evaluate Match accuracy and time accuracy in these conditions. A last improvement would be to allow different number of states and Gaussians per states for HMMs.

8. 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, 94.4% (with 95.2% of phoneme boundaries within 70 ms of hand labelled boundaries) and 96.8% (with 96.1% of phoneme boundaries within 70 ms of hand labelled boundaries). These results suggest that the cost of manual verification and correction of the corpus can be largely reduced. Possible improvements were discussed, among which the use of multiple pronunciations during training, segmentation and labelling error detection.

9. References

F. Béchet. 2001. Lia_phon : un système complet de phonetisation de textes. *Traitement Automatique des Langues*, 42(1):47–68.

Corpatext. 2006. Corpatext 1.02. *www.lexique.org/public.Corpatext.php*.

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

G. D. Fornay. 1973. The Viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–277.

L. Lamel, J. Gauvain, and M. Eskenazi. 1991. BREF, a large vocabulary spoken corpus for french. In *Proc. Eurospeech*.

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

O. Boeffard S. Nefti and T. Moudenc. 2003. Confidence measures for phonetic segmentation of continuous speech. In *Proc. Eurospeech*.

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