Automatic phone segmentation of expressive speech

Laure Charonnat, Gaëlle Vidal and Olivier Boeffard

IRISA - Institut de Recherche en Informatique et Systèmes Aléatoires
Université de Rennes 1, Enssat, Lannion, France
{Laure.Charonnat,Gaëlle.Vidal}@enssat.fr, Olivier.Boeffard@irisa.fr

Abstract

In order to improve the flexibility and the precision of an automatic phone segmentation system for a type of expressive speech, the dubbing into French of fiction movies, we developed both the phonetic labelling process and the alignment process . The automatic labelling system relies on an automatic grapheme-to-phoneme conversion including all the variants of the phonetic chain and on HMM modelling. In this article, we will distinguish three sets of phone models: a set of context independent models, a set of left and right context dependant models and finally a mixing of the two that combines phone and triphone models according to the precision of alignment obtained for each phonetic broad-class. The three models are evaluated on a test corpus. On the one hand we notice a little decrease in the score of phonetic labelling mainly due to pauses insertions, but on the other hand the mixed set of models gives the best results for the score of precision of the alignment.

1. Introduction

Many speech processing systems such as voice conversion, speech recognition or text-to-speech (TTS) synthesis use annotated speech corpora. In the case of a text-to-speech synthesis system based on concatenation of non uniform acoustic units, a speech corpus of about 10 hours is required to obtain a good synthetic speech quality (Kawai and Toda, 2004). These corpora are annotated by a set of phonetic and phonological tags synchronised with the acoustic signal (Torre Toledano et al., 2003). The quality of the produced synthetic signal relies strongly on the phonic segmentation. A manual segmentation process is extremely time consuming. Kawai (Kawai and Toda, 2004) reports that such a segmentation might take up to 130 times the speech duration depending on the language, moreover it requires great skill in acoustic phonetics. Many researchers were interested in automating the segmentation process. The most effective approaches consider HMM modelling (Boeffard et al., 1992) (Ljolje and Riley, 1993) (Brugnara et al., 1993), and a post-processing stage can possibly be defined (Zhao et al., 2005). Concerning the HMM, many studies were interested in optimising modelling and training parameters. Furthermore, many studies make the same assumption considering that the phonetic chain is known and annotated manually.

The contribution that we are suggesting deals with the study of an automatic segmentation system of speech into phones based on HMM modelling [1]. The segmented speech corpus comes from rushes of film dubbing and thus is not primarily dedicated to a TTS system. The phonetic chains are built automatically from an automatic grapheme to phoneme transcription system. One of the main characteristics of this transcription is to take account of the phonological effects of the language, here French, but also of the possible strategies of the speaker (schwas, pauses, etc).

The main goal of this work is to obtain, after segmentation, a corpus of acoustic units making it possible to dub movies using a synthetic voice whose timbre is that of the recorded speaker. This study has two objectives: firstly, to rate the performance of segmentation (labelling and phoneme alignment) on a movie sound track; secondly, we are suggesting an experimental protocol which makes it possible to compare context independent models, context dependent models and an original approach which consists in mixing the two modelling strategies. We decided not to focus here on the various strategies of HMM initialisation (flat initialisation, initialising from speaker independent models or from a manually segmented sub corpus) (Park and Kim, 2007). We have considered the optimal case, which consists in initialising the models by a subset of manually segmented sentences. The various comparisons between context dependent, context independent and mixed models has been done under this optimistic assumption.

The article is organised as follows. Section 2. presents the speech corpus. In section 3., we present the different segmentation systems. Finally the section 4. presents the results obtained.

2. Speech corpus

The speech corpus concerns dubbing into French English voice-over recording of short fantastic stories. The same male speaker reads text prompts in an expressive tone. The corpus lasts 5 hours and 20 minutes. It is composed of 4,995 sentences, with 10 words per sentence on average. After segmentation, the corpus contains 186,215 occurrences of phonemes. The sentences were recorded in a dubbing studio, sampled at 48KHz, and then sub-sampled to 16KHz. The recorded data are presented in the form of blocks of speech separated by long pauses (few seconds). Each block, called speech turn, may contain few words as well as several sentences. The corpus is made up of 1,633 speech turns which last from 1 second to 1 minute and 21 seconds.

Scripts are checked and annotated using the software Transcriber (Barras et al., 1998). The words which deviate from a correct pronunciation and the spelt acronyms are rewritten by taking account of what the speaker said. Deep breathings and long pauses (more than 1 second) are annotated in the text, these annotations are not synchronised with the signal but inserted between two words in the sentence.

[1] This work takes part of the VIVOS project funded by the French National Agency for Research, ANR.

Given the acoustic signal and the associated text, the segmentation system aims to produce a phonetic sequence synchronised with the signal (marks at the beginning and at the end of phonetic segments).
The initial corpus is randomly split into three corpora:

- a learning corpus, corpus A, 70% of the entire corpus, 34,842 words and 131,489 phonemes occurrences, manually segmented, used to estimate HMM modelling parameters and to make decisions for mixing phone and triphone models,
- a validation corpus, corpus B, 12% of the entire corpus, 5,691 words and 21,588 phonemes occurrences, to fix some meta-parameters of the modelling (number of Gaussians)
- a test corpus, corpus C, 18% of the entire corpus, 8,607 words and 32,628 phonemes occurrences, manually segmented, used to evaluate the performance of the segmentation process.

3. Automatic phone segmentation

For each speech turn, a transcription file containing the sequence of phonetic symbols and their synchronisation to the signal is produced by the segmentation system from the acoustic signal and the words sequence. The HMM modelling is built on the HTK tool (Young et al., 2002).

3.1. Grapheme/Phoneme conversion

An automatic phonemic transcription using a set of 36 phonemes of the French language is carried out with the software Lia_phon (Bechet, 2001) which performs a rules-based grapheme-phoneme conversion. Only a few proper names and foreign words are manually transcribed in a small dictionnary (around 600 words). From the Lia_phon outputs, the system produces a graph including optional pauses, breathings and phonological variants without taking punctuation into account. The phonological variants concern liaisons and schwas which are almost all optional, and also the phoneme /ø/ which is optional in some context allowing a good transcription at a high speech rate.
In order to simplify the decoding of a speech turn which could be time and space consuming, *words* in an HTK sense (phonetic sequences delimited by phonemes with no phonological variant) are deduced from the graph (Figure 1). The corpus is made of 69,761 HTK words with an average of 4 phones per HTK word. 58.15% of the words have variants with an average of 4 variants per word.
For each speech turn, the files required by HTK are automatically created from the graph: the Standard Lattice Format file (SLF) of words network hypothesis and the dictionary of word transcriptions with HMM models.

3.2. HMM methodology

An HMM is associated to each phoneme of the phonetic transcript including pauses and breathings.
In this article, we will distinguish three sets of phonemes models: a set of context independent models (HMM-phone), a set of left and right context dependant model (HMM-triphone) and finally a set of mixed models (HMM-mixed) that mixes some phone and triphone models.
In the latter model, a triphone model replaces a phone model when its phonetic class produces a better segmentation score on the learning corpus as described in section 3.3.
For these three modelling strategies, observation vectors are constructed from the first 12 MFCC coefficients augmented with the energy, the first and second derivatives. The observation vectors are normalised and computed every 10 ms on a 32 ms sliding window. All HMM share the same left-to-right topology and have 3 emitting states. For each state, the observation probability is defined by a mixture of gaussians. The number of gaussian components is obtained by the convergence of the likelihood computed on the validation corpus, corpus B. The context dependant models uses 3 gaussian components per state, the context independant models uses 4 gaussians.
Context independent models are initialised on a 34,842 sentences corpus, corpus A. Context dependent models are initialised on context independent models. After the HMM states have been gathered together, a classification tree enables the estimation of missing contextual models for the validation corpus and the test corpus.

3.3. Mixed HMM models

The goal of a segmentation with mixed models is to take advantage of the performance of the context dependent models concerning the phoneme alignment. It is done by excluding from the list of models all the triphone models whose phonetic class leads to poorer or equal results than the ones obtained by context independent models.
In practice, the choice of keeping a triphone model rather than a phone model is based on the analysis of the alignment scores of the corresponding class computed on the learning corpus, corpus A. The phonetic classes used in this study are pauses,voiced plosives, unvoiced plosives, voiced fricatives, unvoiced fricatives, nasal consonants, liquid consonants, semivowels, open oral vowels, close oral vowels, open nasal vowels, and close nasal vowels.
Table 1 presents the segmentation performances of alignment the learning corpus in accordance with phonetic broad classes. For a transition between two phonemes, the first phoneme class is on the Y-axis, the following one on the X-axis. At the intersection of a row and a columnn, for a pair of classes, is found the difference of performance between the segmentations with context independent models and context dependent models; this difference is measured by the number of segmentation marks below the threshold of 20ms. These results show areas where some contextual models are the best (negative values), for example classes [*-plosives], [semivowel-*] and [liquid-*]. In other areas of the table, uncontextual models gives the best results (positive values): [*-fricative] and [pauses] in almost any context. Finally, other areas show disparate results for close phonetic classes, notably vowels. We could notice that in opposition to Toledano (Torre Toledano et al., 2003) remarks, we can find better results of context dependent models on some non stationary phones as plosives.

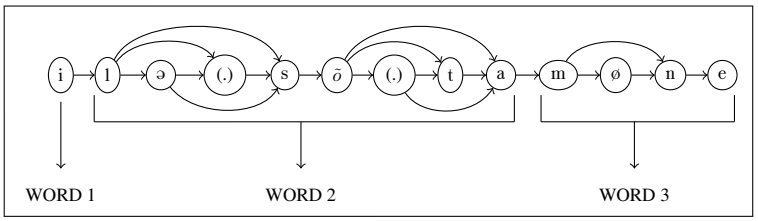

Figure 1: Phonetic transcription of *"ils sont amenés"*. Graph and HTK words.

	P	VP	UVP	VF	UVF	NC	LC	SV	OOV	COV	ONV	CNV
P	7.25	4.24	10.78	14.69	0.43	18.18	2.95	20.00	-0.23	0.36	5.57	3.17
VP	-	-12.74	-12.02	0.93	0.00	0.00	-1.10	-0.94	1.04	-0.83	1.15	0.43
UVP	33.76	3.78	-9.84	0.00	-2.94	-4.49	-2.84	-2.68	-1.59	0.41	-0.34	-0.51
VF	-6.00	-3.82	-1.34	13.69	9.47	-0.09	-2.23	-1.42	-3.18	-1.90	0.20	-1.90
UVF	-4.42	3.68	-0.74	-16.67	1.19	0.00	-3.17	0.11	-0.95	-1.66	-0.84	-0.15
NC	15.39	-14.44	-5.37	0.87	1.75	-12.21	-2.66	-2.03	-2.30	-2.72	-2.02	-1.51
LC	41.80	-2.42	-4.57	1.33	6.96	0.09	-4.19	-5.15	-0.82	-0.72	-0.86	0.64
SV	-0.87	0.00	-3.63	0.00	5.88	8.34	16.67	-	-10.11	-11.97	1.92	2.11
OOV	30.42	-0.41	0.61	-2.67	3.19	-0.26	-1.20	-1.77	-5.63	2.30	-3.55	-8.44
COV	17.87	-0.86	-0.45	-0.50	2.65	-0.34	-3.63	-2.13	-12.69	-2.27	-7.67	4.94
ONV	14.42	-1.71	-6.73	1.19	2.10	-1.96	-3.22	0.00	-13.10	1.18	0.00	3.58
CNV	28.02	-1.95	-2.24	-3.78	1.35	-1.76	-1.96	0.00	13.80	-5.22	16.66	8.70

Table 1: This table presents differences of percentage of correct segmentation below a threshold of 20 ms between context independent models and context dependent models. Phonetic classes are defined in the following way: P: Pauses, VP: Voiced Plosives, UVP: UnVoiced Plosives, VF: Voiced Fricatives, UVF: UnVoiced Fricatives, NC: Nasal Consonants, LC: Liquid Consonants, SV: SemiVowels, OOV: Open Oral Vowel, COV: Close Oral Vowel, ONV: Open Nasal Vowel, CNV: Close Nasal Vowel. The character [-] represents a non-existing transition.

4. Results and discussion

Two types of results computed on the test corpus, corpus C, are presented. Section 4.1. presents an evaluation of the phonetic decoding and section 4.2. presents the scores of the alignment on the acoustical signal obtained by *HMM-phone*, *HMM-triphone* and *HMM-mixed* systems. The selection of the triphone models defining the HMM-mixed system leads to a proportion of 65.67% of phone models and 33.34% of triphone models for this test corpus.

4.1. Phonetic decoding

The conformity of a phonetic sequence is measured by comparing phone sequences defined by an automatic and a manual segmentation.

Table 2 presents percentages of missing labels (Eli), inserted phonetic (Ins) and substitutions (Sub) measured on the test corpus.

	Eli	Ins	Sub
HMM-phone	0.32% [±0.06%]	1.01% [±0.11%]	3.92% [±0.21%]
HMM-triphone	0.22% [±0.05%]	0.90% [±0.10%]	3.99% [±0.21%]
HMM-mixed	0.26% [±0.06%]	1.30% [±0.12%]	3.99% [±0.21%]

Table 2: Percentage of error on phonetic labels on the test sets with 95% confidence intervals.

Whichever the model type, only a few phones are missing or inserted in the automatic phonetic sequence, basically schwas and pauses. For example, concerning the HMM-mixed models, errors on schwas represent 53% of the total number of elisions and 32% of the number of insertions. Errors on auses represent 22% of all the elisions and 60% of the insertions.

Substitutions are mostly due to an inversion between open and closed vowels. For example, concerning the HMM-mixed model, 35% of the substitutions are due to a replacement of /e/ by a /ɛ/ in the phonetic sequence, and 19% are some /ə/ replaced by /ø/. This errors can be imputated to the speaker's pronunciation especially concerning /e/ and /ɛ/.

The percentage disagreement defined as

$$(Elisions + Insertions + Substitutions)/N * 100\%$$

with N the total number of phones in the manual transcription (Van Bael et al., 2007) is 5.25% for the context dependent models, 5.11% for the context independent models and 5.55% for the mixed models. Theses results are high as compared to human inter-labeller disagreement scores reported by Van Bael (Van Bael et al., 2007) that is to say 6.9-5.6% for German read speech and 6.2-3.7% for Dutch read speech.

These good results can be attributed to the good accuracy of the system and the use of a phonological graph but also to the clear pronuncition of the speaker and the kind of speech studied here. Indeed as shown by Toledano in (Torre Toledano et al., 2005) best results in phonetic decoding are obtained on formal speech which is the kind of speech this study refers to.

4.2. label alignments

The precision of the alignment is measured comparing manual and automatic phone transitions on well recognised phonetic labels.

	≤ 10 **ms**	≤ 20 **ms**	≤ 30 **ms**
HMM-phone	74.98% [$\pm$0.48%]	93.56% [$\pm$0.27%]	97.55% [$\pm$0.17%]
HMM-triphone	77.00% [$\pm$0.47%]	93.51% [$\pm$0.27%]	97.17% [$\pm$0.18%]
HMM-mixed	78.57% [$\pm$0.46%]	94.84% [$\pm$0.24%]	98.50% [$\pm$0.16%]

Table 3: Proportion of alignment marks below different thresholds (10 ms, 20 ms, 30 ms) comparing automatic and manual segmentation.

The mean lag of the marks of well recognised phonetic labels is 9.41 ms $\pm$0.20 for the context independent models, 9.86 ms $\pm$0.30 for the context dependent models and 8.54 ms $\pm$0.19 for the mixed models.

It is observed that mixed models are more precise than phone or triphone models and also that triphone models have better precision than phone models for small tolerance ($<$ 10 ms). The mixed models take advantage of context dependent model performance concerning phones which strongly relies on their neighbourhoods. The replacement of a triphone model by a phone model can have a longer term influence. Indeed, during the Viterbi decoding, the choice of a model influences its closest neighbour; this can lead, locally, to a less precise segmentation.

5. Conclusion

This article presents an analysis of three automatic phone segmentation systems applied to an expressive speech corpus coming from the dubbing of a movie. The first of these automatic segmentation systems is based on context independent models. The second uses context dependent models and the last uses a mix of the two. The use of the mixed models is motivated by the fact that the quality of the segmentation of some co-occurring phonetic classes depends on whether the model is context dependent or not. The HMM are initialized from a manual segmentation of the learning corpus and the number of mixture gaussian components are defined using the validation corpus. The learning corpus is also used to choose, for each triphone which type of models is to be used in the mixed models systeme according to the performances of its phonetic class in context.

For the evaluation on the test or the validation corpus, only the text of the sentences is known. The phonemic sequence is determined automatically and the phonetic labels are obtained by Viterbi decoding.

On the test corpus and for the three systems, less than 6% of phonetic labels are wrong and more than 93% of the segmentation labels are closer than 20 ms to the manual segmentation labels. The use of mixed models increases the alignement precision by 1.33% for a 20 ms threshold compared to context dependent models and by 1.27% compared to context independent models. These results could be improved in a future work, especially phonetic decoding by adding a post-processing stage focusing on pauses and open/close vowels discriminations. Further investigations will concern more expressive speech including non-linguistic events made by the speaker such as laughter, noises, etc.

6. References

Barras, C., E. Geoffrois, Z. Wu, and M. Liberman, 1998. Transcriber: a free tool for segmenting, labeling and transcribing speech. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*. Arcueil, France, and Philadelphia, USA.

Bechet, F., 2001. Liaphon - un systeme complet de phonetisation de textes. *Traitement Automatique des Langues (T.A.L.) edition Hermes*, 42(1).

Boeffard, O., L. Miclet, and S. White, 1992. Automatic generation of optimal unit dictionaries for text-to-speech synthesis. In *proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP'92)*. Banff, Alberta, Canada.

Brugnara, F., D. Falavigna, and M. Omologos, 1993. Automatic segmentation and labeling of speech based on hidden markov models. *Speech Communication*, 12:357–370.

Kawai, H. and T. Toda, 2004. An evaluation of automatic phone segmentation for concatenative speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing 2004 (ICASSP apos;04)*. Montréal, Québec, Canada.

Ljolje, A. and M.D. Riley, 1993. Automatic segmentation of speech for tts. In *proceedings of the third European Conference on Speech Communication and Technology (EUROSPEECH'93)*. Berlin, Germany.

Park, S.S. and N.S. Kim, 2007. On using multiple models for automatic speech segmentation. *Audio, Speech and Language Processing, IEEE Transactions on*, 15(8):2202–2212.

Torre Toledano, D., A. Hernandez Gomez, and L. Villarrubia Grande, 2003. Automatic phonetic segmentation. *IEEE Transactions on Speech and Audio Processing*, 11(6):617–625.

Torre Toledano, D., A. Moreno Sandoval, L. Colas Pasamontes, and J. Garrido Salas, 2005. Acoustic-phonetic decoding of different types of spontaneous speech in spanish. In *proceedings of Disfluency in Spontaneous Speech Workshop (DiSS'05)*. Aix-en-Provence, France.

Van Bael, C., L. Boves, H. van den Heuvel, and H. Strik, 2007. Automatic phonetic transcription of large speech corpora. *Computer Speech and Langage*, 21:652–668.

Young, S., G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, 2002. In *The HTK Book (for HTK Version 3.2)*. Cambridge, U.K.

Zhao, Y., L. Wang, M. Chu, F. K. Soong, and Z. Cao, 2005. Refining phoneme segmentations using speaker-adaptative context dependent boundary models. In *Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech)*. Lisbon, Portugal.