Impact of frame rate on automatic speech-text alignment for corpus-based phonetic studies

Katarina Bartkova, Denis Jouvet

IMPACT OF FRAME RATE ON AUTOMATIC SPEECH-TEXT ALIGNMENT FOR CORPUS-BASED PHONETIC STUDIES

Katarina Bartkova[1], Denis Jouvet[2]

[1] ATILF - Analyse et Traitement Informatique de la Langue Française
Université de Lorraine, ATILF, UMR 7118, Nancy, F-54063, France

[2] Speech Group, LORIA
Inria, Villers-lès-Nancy, F-54600, France
Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
CNRS, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France

katarina.bartkova@atilf.fr, denis.jouvet@loria.fr

ABSTRACT

Phonetic segmentation is the basis for many phonetic and linguistic studies. As manual segmentation is a lengthy and tedious task, automatic procedures have been developed over the years. They rely on acoustic Hidden Markov Models. Many studies have been conducted, and refinements developed for corpus based speech synthesis, where the technology is mainly used in a speaker-dependent context and applied on good quality speech signals. In a different research direction, automatic speech-text alignment is also used for phonetic and linguistic studies on large speech corpora. In this case, speaker independent acoustic models are mandatory, and the speech quality may not be so good. The speech models rely on 10 ms shift between acoustic frames, and their topology leads to strong minimum duration constraints. This paper focuses on the acoustic analysis frame rate, and gives a first insight on the impact of the frame rate on corpus-based phonetic studies.

Keywords: Automatic speech-text alignment, frame rate, pronunciation variants.

1. INTRODUCTION

Phonetic segmentation, i.e., segmentation of the speech signal into phones and words, is the basis for many phonetic and linguistic studies, as well as for developing corpus-based speech synthesis systems. As manual segmentation is a lengthy and tedious work, automatic segmentation procedures have been developed over the year. Although some studies were carried out for speech-text alignment on long recordings [14], most of the automatic speech-text alignment systems deals with speech segments that are the size of a sentence (e.g., [28], [12], [4]).

Automatic speech-text alignment systems usually rely on speech recognition technologies, and more precisely on hidden Markov models (HMM), with frame features computed every 10 ms. However the context-dependent phone modelling that provides the best performance in speech recognition is not necessarily the most efficient with respect to boundary accuracy for speech-text alignment; context independent phone models usually lead to more accurate boundaries [20]. Viterbi-based alignment has been compared to forward-backward procedures [8] and boundary statistical corrections were proposed for context-dependent-based modelling [30]. Impact of the model topology [27] and of segmentation constrained training [15] were also investigated.

Many segmentation procedures were refined and evaluated in the framework of corpus-based text-to-speech synthesis (e.g., [19]). Although speech recognition technology typically computes 100 frames per second, that is a 10 ms frame shift, higher frame rates corresponding to 3 ms [30] , 4 ms [3] or 5 ms [25] frame shift have been used for speech segmentation for improving the boundary precision. It should be noted that for concatenative speech synthesis speaker adapted or speaker dependent models are used, and that the speech signal is of very good quality. Boundary refinement post-processing was also proposed using other features or techniques targeted towards the detection of transitions [30], possibly through multi-layer

perceptron [24] or support vector machine [22] approaches. The use of multiple features [25], of multiple models [21] and of multiple systems [17] were also investigated.

A different research direction consists in using automatic speech-text alignment for conducting phonetic and linguistic studies on large speech corpora [1]. This includes the study of the schwa and of liaisons [8], [7], [5], as well as the study of pronunciation variants [2] and the analysis of other phenomena [23], [26]. In these approaches speaker-independent models are required, and the speech signal is not always of good quality. These studies were conducted using the standard frame rate (that is a shift of 10 ms between frames). Moreover, the model topology leads to a minimum duration of three frames, i.e. 30 ms, for each phone segment.

Because such minimum phone duration constraint impacts on the phone segmentation, this paper focuses on a first analysis of the impact of the frame rate on corpus-based phonetic studies.

The paper is organized as follows. Section 2 presents the speech corpora used, and section 3 details the automatic speech-text alignment process. Section 4 presents an insight on the impact of the frame rate for phonetic studies. A conclusion ends the paper.

2. SPEECH CORPORA

The speech corpora used in the experiments come from the ESTER2 [11] and the ETAPE [13] evaluation campaigns, as well as from the EPAC [10], [9] project.

The ESTER2 and EPAC data are French broadcast news collected from various radio channels. They contain mainly prepared speech (speech from the journalists). A large part of the data is of studio quality, though some parts are of telephone quality. On the opposite, the ETAPE data corresponds to debates collected from various radio and TV channels. Thus this corresponds mainly to spontaneous speech.

Only the train subsets of these corpora are used in the experiments reported in this paper. This amounts to about 280 hours of signal for which a manual orthographic transcription, at the word level, is available.

3. AUTOMATIC SPEECH-TEXT ALIGNMENT

The transcribed data is used for training the acoustic model parameters, and for speech-text phonetic alignments. The Sphinx speech recognition toolkit [29] is used in the reported experiments.

3.1. Training the speech models

In order to train the acoustic models, pronunciation variants of the words of the training set are generated. Whenever possible, they are extracted from available lexicons (BDLEX [6] and in-house lexicons). For words not present in these lexicons, the pronunciation variants are obtained automatically using joint multigram models (JMM) and conditional random field (CRF) based grapheme-to-phoneme converters, similar to what is described in [16]. On average, there are 2.25 pronunciations variants per word in the training lexicon. Most of the pronunciation variants come from the mute ‘e’ (schwa /ə/ which can be pronounced or not at the end of many words, or in internal position in some French words), and from the liaisons (i.e. introduction of a liaison consonant which may be pronounced when the following word starts by a vowel).

For each model, the training is carried out in two successive passes. The first training pass relies on a default pronunciation variant for each word. The resulting acoustic models are used to automatically align the training data in order to find the best matching pronunciation variant of each word in each utterance of the training data. This alignment is then used for the second training pass.

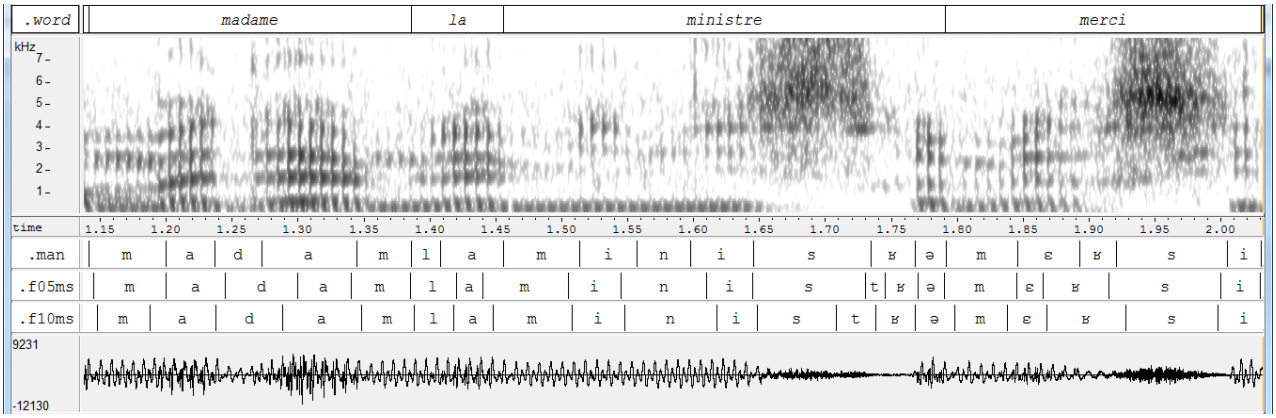
Context-dependent models are estimated for each phone. 4500 shared densities are estimated for each set of acoustic models, each having 64 Gaussian components.

3.3. Frame rate and speech-text alignment

The acoustic analysis of the training data is performed two times, with respectively 10 ms and 5 ms frame shifts; standard Mel frequency cepstral coefficients (MFCC) features are computed. The 10 ms frame shift is the standard value usually used for speech recognition and speech-text alignment. The 5 ms frame shift leads to two times more frames.

The hidden Markov models associated to each phone have three emitted states, without skip. This leads to a minimum duration of three frames for

Figure 1: Example of manual and automatic phone segmentation ("*.man*" indicates the manual segmentation, "*.f05ms*" the automatic segmentation using 5 ms frame shift, and "*.f10ms*" the automatic segmentation using 10 ms frame shift).



each phone segment; thus a minimum of 30 ms for the 10 ms frame shift case, and a minimum of 15 ms for the 5 ms frame shift case.

It should be noted that for the 5 ms frame shift, the computation of the temporal derivatives is adjusted to consider the same temporal window as the standard computation of the derivatives for the standard 10 ms frame shift.

The final acoustic models obtained in the second training pass are used for a last speech–text alignment, which is discussed hereafter.

4. IMPACT OF FRAME RATE

The speech-text alignment fails for a few utterances: out of the almost 300 000 utterances of the train corpora, 2058 utterances failed to be aligned with the 10 ms frame shift, and only 1380 failed to be aligned with the 5 ms frame shift. The shorter phone minimum duration reduces the constraints on the speech-text alignment process, and this probably explains this difference.

This section gives an insight of the impact of the frame shift (and thus associated minimum phone duration) on some phonetic aspects.

4.1. Impact on phone boundaries

Figure 1 above displays an example of speech alignment. Besides the orthographic transcription and the spectrogram (in top panels) and the waveform (bottom panel), the first line displays the manual segmentation made by a phonetician (panel "*.man*"), the second one displays the automatic alignment using the 5 ms frame shift (panel "*.f05ms*") and the third one displays the automatic alignment achieved with the 10 ms frame shift (panel "*.f10ms*"). The 10 ms frame shift is the default value used in speech recognition systems.

The manual segmentation was carried out by a phonetician from scratch to avoid the usual bias of the manual "verification and correction" process, where only the boundaries which are notably wrong are corrected.

The French sentence of this example is "…*Madame la Ministre merci*…" ("…Madame Minister thanks…") pronounced in a rather rapid speaking mode. The phonetician did not observe any presence of a /t/ at the end of the word "*Ministre*", but just a short /ʁ/ and a short schwa /ə/. As the pronunciation variant without /t/ is not present in the pronunciation lexicon, the automatic alignments found, in both cases, that the pronunciation variant providing the best match is /m i n i t ʁ ə/. However, with the 5 ms frame shift, the part /t ʁ ə/ corresponds to three short segments (and the /t ʁ/ segments almost corresponds to the /ʁ/ segment of the manual annotation), whereas for the 10 ms frame shift, the 30 ms phone minimum constraint force the /t/ to a wrong temporal position (where it overlaps with the actual /s/ sound of the manual segmentation).

This example shows that having a shorter phone minimum duration constraint helps when dealing with rapid speaking rate, although it is sometime difficult to decide in fast speaking rate if a sound is reduced (in duration) or is discarded by the speaker.

As the speech-text alignment is carried out using context-dependent phone models, we will not discuss in details the exact position of the boundaries with respect to the manual

segmentation. For such a discussion, the alignment should be carried out using context-independent phone models, or some boundary corrections or refinements should be applied, as discussed in the introduction. However context-dependent phone models takes better into account the transition between the phones, and thus, should lead to better performance for selecting the most relevant pronunciation variants.

4.2. Impact on pronunciation variants statistics

This second analysis presents and discusses some statistics on the frequency of the pronunciation variants when estimated from the 5 ms frame shift based alignment and from the 10 ms frame shift base alignment.

Table 1: Frequency of pronunciation variants for a few words, estimated from the speech-text alignments using the 5 ms and the 10 ms frame shift.

Word	Variant	Variant frequency	
		05 ms	10 ms
de (of, from)	/d ə/	78 %	77 %
	/d/	22 %	23 %
que (that, which)	/k ə/	80 %	78 %
	/k/	20 %	22 %
une (a, one)	/y n/	75 %	80 %
	/y n ə/	25 %	20 %
dire (say)	/d i ʁ/	86 %	92 %
	/d i ʁ ə/	14 %	8 %
petit (small)	/p t i/	43 %	59 %
	/p ə t i/	40 %	27 %
	/p ə t i t/	11 %	8 %
	/p t i t/	6 %	6 %

One of the main pronunciation variants in the lexicon come from the mute 'e' (schwa /ə/) which can be pronounced or not. Table 1, reports the frequency of the pronunciation variants for these two frame shift automatic alignments. For the final schwa, i.e. the first four lines, the frequency of the variant including the final schwa is somewhat higher in the 5 ms frame shift alignments, from 1% or 2% more for the words "*de*" and "*que*" up to 6% more for the word "*dire*".

For the last line, which concerns the word "*petit*", a similar phenomenon is observed. The pronunciation variants that include the schwa are much more frequent in the alignment realized using the 5 ms frame shift analysis, than in the alignments resulting from the 10 ms frame shift analysis.

5. CONCLUSION

In this paper we have started investigating the impact of the frame rate used in the acoustic analysis when applied for automatic speech-text alignment. Using a higher frame rate than the usual 100 frames per second standard feature analysis, reduces the 30 ms minimum phone duration constraint (which results from the three emitting states of the hidden Markov models used for each phone), down to 15 ms when a 5 ms frame shift is used.

This reduction of the phone minimum duration constraint leads to differences in the phone segmentation, especially in fast speaking rate, as well as in differences in the statistics of the frequency of the pronunciation variants measured on a large speech corpus.

These two results shows that the frame shift impact on corpus-based phonetic analysis.

Future work will investigate a more refined comparison with respect to the estimated speaking rate. In [18] an analysis of the frequency of occurrences of the final schwa showed that the final schwa (French mute 'e') was less and less frequent when the speaking rate increases. A similar study considering the 5 and 10 ms frame shift alignments should then provide interesting statistics.

7. REFERENCES

[1] Adda-Decker, M. 2006. De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux. *Proc. JEP'2006*, Dinard, France, 389-400.

[2] Adda-Decker, M., Lamel, L. 2000. Systèmes d'alignement automatique et études de variantes de prononciation. *Proc. JEP'2000*, Aussois, France, 189-192.

[3] Adell, J., Bonafonte, A., Gómez, J. A., Castro, M. J. 2005. Comparative study of Automatic Phone Segmentation methods for TTS. *Proc. ICASSP'2005*, Philadelphia, USA, 309-312.

[4] Bigi, B., Hirst, D., 2012. SPeech Phonetization Alignment and Syllabification (SPPAS): a tool for the automatic analysis of speech prosody. *Proc. Speech Prosody*, Shanghai, China, 1-4.

[5] Bürki, A., Gendrot, C., Gravier, G., Linarès, G., Fougeron, C. 2008. Alignement automatique et analyse phonétique: comparaison de différents systèmes pour l'analyse du schwa. *Traitement Automatique des Langues*, 49(3), 165-197.

[6] de Calmès, M., Pérennou, G. BDLEX : a Lexicon for Spoken and Written French. *Proc. LREC'1998*, Grenada, Spain. 1998, 1129-1136.

[7] De Mareüil, P. B., Adda-Decker, M., Gendner, V. 2003. Liaisons in French: a corpus-based study using morpho-syntactic information. *Proc. ICPhS'2003*, Barcelona, Spain.

[8] Demuynck, K., Laureys, T. 2002. A comparison of different approaches to automatic speech segmentation. *Proc. Text, Speech and Dialogue*, Brno, Czech Republic, 277-284.

[9] EPAC Corpus: Orthographic transcriptions, ELRA catalogue (http://catalog.elra.info), ref. ELRA-S0305.

[10] Estève, Y., Bazillon, T., Antoine, J.-Y., Béchet, F., Farinas, J. 2010. The EPAC corpus: Manual and automatic annotations of conversational speech in French broadcast news. *Proc. LREC'2010*, Valetta, Malta.

[11] Galliano, S., Gravier, G., Chaubard, L. 2009. The Ester 2 evaluation campaign for rich transcription of French broadcasts. *Proc. INTERSPEECH'2009*, Brighton, UK, 2583-2586.

[12] Goldman, J. P. 2011. EasyAlign: an automatic phonetic alignment tool under Praat.

[13] Gravier, G., Adda, G., Paulsson, N., Carré, M., Giraudel, A., Galibert, O. 2012. The ETAPE corpus for the evaluation of speech-based TV content processing in the French language, *Proc. LREC'2012*, Istanbul, Turkey.

[14] Hoffmann, S., Pfister, B. 2013. Text-to-speech alignment of long recordings using universal phone models. *Proc. INTERSPEECH'2013*, Lyon, France, 1520-1524.

[15] Huggins-Daines, D., Rudnicky, A. I. 2006. *A Constrained Baum-Welch Algorithm for Improved Phoneme Segmentation and Efficient Training*, CMU report.

[16] Illina, I., Fohr, D., Jouvet, D. 2011. Grapheme-to-Phoneme Conversion using Conditional Random Fields. *Proc. INTERSPEECH'2011*, Florence, Italy.

[17] Jarifi, S., Pastor, D., Rosec, O. 2008. A fusion approach for automatic speech segmentation of large corpora with application to speech synthesis. *Speech Communication*, 50(1), 67-80.

[18] Jouvet, D., Fohr, D., Illina, I. 2010. Detailed pronunciation variant modeling for speech transcription. *Proc. INTERSPEECH'2010*, Makuhari, Japan.

[19] Kawai, H., Toda, T. 2004. An evaluation of automatic phone segmentation for concatenative speech synthesis. Proc. *ICASSP'2004*, Montreal, CA, vol. I, 677-680.

[20] Kessens, J. M., Strik, H. 2004. On automatic phonetic transcription quality: lower word error rates do not guarantee better transcriptions. *Computer Speech & Language*, 18(2), 123-141.

[21] Kominek, J., Black, A. W. 2004. A family-of-models approach to HMM-based segmentation for unit selection speech synthesis. *Proc. INTERSPEECH'2004*, Jeju Island, Korea.

[22] Kuo, J. W., Lo, H. Y., Wang, H. M. 2007. Improved HMM/SVM methods for automatic phoneme segmentation. *Proc. INTERSPEECH'2007*, Antwerp, Belgique, 2057-2060.

[23] Kuperman, V., Pluymaekers, M., Ernestus, M., Baayen, H. 2007. Morphological predictability and acoustic duration of interfixes in Dutch compounds. *Journal of the Acoustical Society of America*, 121(4), 2261-2271.

[24] Lee, K. S. 2006. MLP-based phone boundary refining for a TTS database. *IEEE Trans. on Audio, Speech, and Language Processing*, 14(3), 981-989.

[25] Mporas, I., Ganchev, T., Fakotakis, N. 2010. Speech segmentation using regression fusion of boundary predictions. *Computer Speech & Language*, 24(2), 273-288.

[26] Nakamura, M., Iwano, K., Furui, S. 2008. Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech & Language*, 22(2), 171-184.

[27] Ogbureke, K. U., Carson-Berndsen, J. 2009. Improving initial boundary estimation for HMM-based automatic phonetic segmentation. *Proc. INTERSPEECH'2009*, Brighton, UK, pp. 884-887.

[28] Sjölander, K. 2003. An HMM-based system for automatic segmentation and alignment of speech. *Proc. of Fonetik*, Lövånger, Sweden, 93-96.

[29] Sphinx. [Online]: http://cmusphinx.sourceforge .net/, 2011.

[30] Toledano, D. T., Gómez, L. A. H., Grande, L. V. 2003. Automatic phonetic segmentation. *IEEE Trans. on Speech and Audio Processing*, 11(6), 617-625.