Acoustical Frame Rate and Pronunciation Variant Statistics

Denis Jouvet, Katarina Bartkova

Acoustical Frame Rate and Pronunciation Variant Statistics

Denis Jouvet[1] and Katarina Bartkova[2]

[1] 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
denis.jouvet@inria.fr
[2] ATILF - Analyse et Traitement Informatique de la Langue Française
44 Av De La Libération, BP 30687, 54063 Nancy Cedex, France
katarina.bartkova@atilf.fr

Abstract. Speech technology enables computing statistics on word pronunciation variants as well as investigating various phonetic phenomena. This is achieved through a forced alignment of large amounts of speech signals with their possible pronunciations variants. Such alignments are usually performed using a 10 ms frame shift acoustical analysis. Therefore, the three emitting state structure of conventional acoustic hidden Markov models introduces a minimum duration constraint of 30 ms for each phone segment. This constraint is not critical at low speaking rates, but may introduce artefacts at high speaking rates. Thus, this paper investigates the impact of the acoustical frame rate on corpus-based phonetic statistics. Statistics on pronunciation variants obtained with a shorter frame shift (5 ms) are compared to the statistics resulting from the standard 10 ms frame shift. Statistics are computed on a large speech corpus of more than 3 million running words, and are analyzed with respect to the estimated local speaking rate. Results exhibit some discrepancies between the two sets of statistics, in particular for high speaking rates where the usual acoustic analysis frame shift of 10 ms leads to an under-estimation of the frequency of the longest pronunciation variants.

Keywords: speech modeling, speech-text alignment, acoustical frame rate, corpus-based phonetic statistics

1 Introduction

Many phonetic studies rely on a segmentation of the speech signal into words and phones. Manual segmentation, especially at the phone level, is a lengthy and tedious task. Moreover the agreement between annotators is not perfect even with respect to the existence of some phone segments [26] in spontaneous speech. Another approach consists in relying on automatic segmentation at the phone and at the word levels of large amounts of speech data. In such an approach the

manual transcription of the speech signal into words is still required, but this is a much easier task than the phonetic segmentation itself. Knowing the sequence of words corresponding to a speech segment, all the possible pronunciations into sequences of phones are derived, and automatic alignment tools find the sequence of phones (among all the possible phone sequences corresponding to the different pronunciation variants) that best matches with the speech signal.

Speech-text alignments are typically performed on speech segments that are the size of a sentence (e.g., [27, 13, 4]) using hidden Markov models (HMM) with feature vectors computed every 10 ms. Although context dependent phone models provide the best performance in speech recognition because of a better modeling of the co-articulation between adjacent sounds, more accurate boundaries are usually obtained with context independent phone models [20]. Boundary statistical corrections were proposed for context-dependent based modeling [29], and segmentation constrained training [15] was investigated, as well as the impact of the model topology [25].

Many phone segmentation procedures were improved and evaluated for corpus-based speech synthesis (e.g. [19]). In this context, higher acoustic analysis frame rates corresponding to 3 ms [29], 4 ms [3] or 5 ms [23] frame shifts are used for improving the phone boundary precision. Moreover, boundary refinement post-processing was also proposed, for example, using other features or techniques targeted towards the detection of transitions [29]. Other proposed approaches consist in using multiple features [23], multiple models [21] or multiple systems [17]. It should be noted that all these approaches have been developed for corpus-based speech synthesis, so they are dealing with good quality speech signal, well-articulated speech, and the sequence of phones corresponding to each sentence is assumed to be known (because of the controlled recording that is carried out for such corpora).

Another research direction relates to the use of 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 [9, 8, 5], as well as the study of pronunciation variants [2] and the analysis of other phenomena [22, 24]. In these approaches speaker-independent models are required, and the acoustic models typically rely on 10 ms frame shift between adjacent feature vectors. Consequently, the three emitting states of the acoustic models lead to a minimal duration of three frames (i.e., 30 ms) for each phone segment. As such minimum phone duration is a constraint for the phone segmentation process, this paper focuses on the analysis of the impact of the acoustic analysis frame rate on corpus-based phonetic statistics. Several pronunciation phenomena are studied and analyzed with respect to the local speaking rate. The mute "e" is particularly studied as the phonetic realization or the elision of the corresponding sound (/ə/) is one main adjustment variable of the speaking rate in French (further comments are given in Section 4.1). Another aspect studied is the pronunciation of some consonantal clusters in word final position (detailed in Section 4.2).

The paper is organized as follows. Section 2 presents the speech data and the modeling used while Section 3 details the speech-text alignment process. Then,

Section 4 analyzes the frequency of some pronunciation variants with respect to the local speaking rate, using speech-text alignments obtained with 5 ms and 10 ms frame shifts. A conclusion ends the paper.

2 Speech data and modeling

The speech corpora used in the experiments come from the ESTER2 [12] and the ETAPE [14] evaluation campaigns, and from the EPAC [11, 10] 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. The ETAPE data corresponds to debates collected from various radio and TV channels. Thus this corresponds mainly to spontaneous speech. Only the training subsets of these corpora are used in the experiments reported in this paper. This amounts to almost 300 hours of speech signal for which a manual orthographic transcription, at the word level, is available.

The speech material was analyzed by computing 13 Mel frequency cepstral coefficients (MFCC) per frame. The whole data set was analyzed two times, once with the standard 10 ms frame shift, and once with the reduced 5 ms frame shift. First and second order temporal derivatives were then added to the static coefficients to produce a 39 coefficient vector. For the 5 ms frame shift, the indexes of the frames involved in the computation of the temporal derivatives were modified in order to correspond to the same temporal window as in the 10 ms frame shift case.

Using the conventional modeling approach, each phone was modeled by a three emitting state hidden Markov model. The training process involved several steps. First, using the standard pronunciation of each word, a first model was trained, and then used to realign the training data in order to associate with each speech segment the sequence of estimated pronunciation variants of the words. A second model was then trained from these alignments, and the training data were re-aligned a second time using this second model. Finally a third, more detailed, model was trained (7500 shared densities – senones), and was then used for determining the speech-text alignments that are later analyzed in the paper. The trained acoustic models have 64 Gaussian components per density.

This training procedure was applied for each frame shift (5 and 10 ms frame shifts), and for each pronunciation lexicon (see details in Section 3.1), using the Sphinx toolkit [28].

3 Speech-text alignment

All the training data were aligned with the trained acoustic models, thus providing the phone and word segmentations.

3.1 Lexicon and pronunciation variants

The lexicon contains more than 60,000 words. Whenever possible the pronunciation variants of the words were extracted from available lexicons (BDLEX [7] and in-house lexicons). For the words not present in these lexicons, the pronunciation variants were obtained automatically using joint multigram models (JMM) and conditional random field (CRF) based grapheme-to-phoneme converters [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). This corresponds to the standard lexicon which is used in the experiments reported later on in Sections 3.2 and 4.1.

An extended lexicon was created for the last set of experiments (in Section 4.2). It corresponds to the standard lexicon to which extra pronunciation variants were added for words that end by a cluster /plosive liquid/ such as /t ʁ/ in final position in the word "*ministre*" (minister). In fact, in order to analyze the pronunciation of such final clusters, all the possible pronunciation variants were generated, considering the phonetic realization or the elision of any of the corresponding phonemes, as well as the pronunciation of a final schwa. Hence for the word "*ministre*" we get eight pronunciations variants, as shown in Table 1.

Table 1. The eight pronunciation variants for the word "*ministre*" ("minister") in the extended pronunciation lexicon.

/m i n i s t ʁ ə/	[+t][+ʁ][+ə]	/m i n i s ʁ ə/	[-t][+ʁ][+ə]
/m i n i s t ʁ/	[+t][+ʁ][-ə]	/m i n i s ʁ/	[-t][+ʁ][-ə]
/m i n i s t ə/	[+t][-ʁ][+ə]	/m i n i s ə/	[-t][-ʁ][+ə]
/m i n i s t/	[+t][-ʁ][-ə]	/m i n i s/	[-t][-ʁ][-ə]

3.2 Example of phone segmentation

Figure 1 displays an example of speech alignment. The panel "man" displays the manual segmentation, and the panels ".f05ms" and ".f10ms" display the automatic segmentations achieved with the standard pronunciation lexicon and, respectively, the 5 and 10 ms frame shifts. The French sentence of this example is "*... Madame la Ministre merci ...*" ("... Madame Minister thanks ...") pronounced in a rather rapid speaking rate. Our expert phonetician has not observed 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 standard pronunciation lexicon used, the automatic alignments

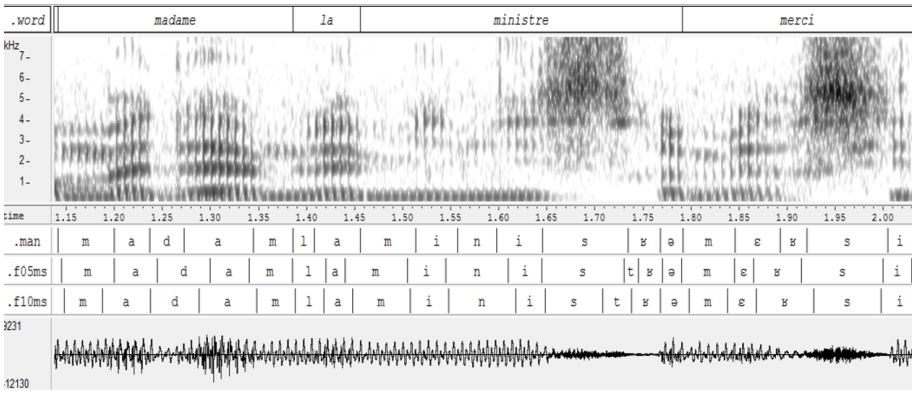

Fig. 1. Example of manual and automatic phone segmentations (".man" indicates the manual segmentation, ".f05ms" the automatic segmentation using 5 ms frame shift, and ".f10ms" the automatic segmentation using 10 ms frame shift). The speech segment corresponds to "*... Madame la Ministre merci ...*" ("... Madame Minister thanks ...") pronounced in a rather rapid speaking rate.

found, in both cases, that the pronunciation variant providing the best match is /m i n i s t ʁ ə/. However, with the 5 ms frame shift, the part /t ʁ ə/ corresponds to three short segments (and the /t ʁ/ segment almost correspond 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.

4 Impact of frame rate on statistics

This section analyzes, with respect to the local speaking rate, statistics on pronunciation variants estimated using 5 and 10 ms frame shifts. The local speaking rate is computed for each word using a local window of at most seven words (three words before and three words after the current word), similar to what was done in [18]. A smaller number of words may be considered if a long pause (more than 100 ms) is present in this window. In the reported statistics, the local speaking rate is expressed in phones per second.

4.1 Mute "*e*"

Figure 2 shows the frequency of the variants corresponding to the pronunciation of the mute "*e*" in function words followed by a word starting with a

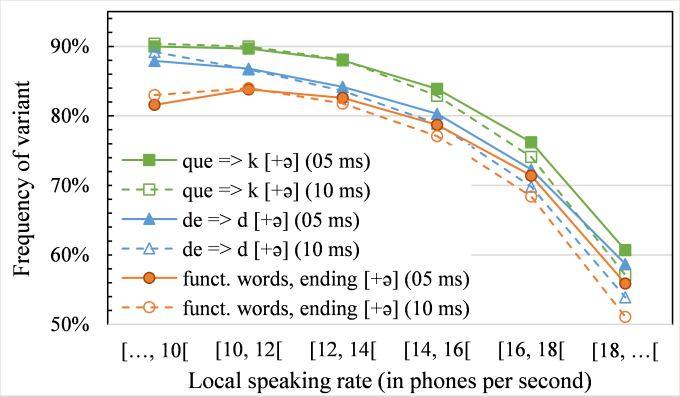

Fig. 2. Frequency of pronunciation of the final schwa, estimated using 5 or 10 ms frame shifts, in short words such as "*que*" (that) and "*de*" (of), and average over all similar monosyllabic function words.

consonant. The frequency of these pronunciation variants are displayed for the function words "*que*" (meaning "that", 150,000 occurrences) and "*de*" (meaning "of", 38,000 occurrences). The last curves correspond to the frequency of pronunciation of the mute "*e*" estimated over all similar function words (340,000 occurrences).

For the word "*que*", at low speaking rate, in 90% of the cases the final schwa is pronounced. The frequency of pronunciation of the schwa gets lower as the speaking rate increases. The figure shows that at low speaking rate, the 5 and 10 ms frame shifts lead to rather similar frequency values. However, as the speaking rate increases, the difference between the statistics estimated with the 5 and 10 ms frame shifts gets larger, the 5 ms frame shift leading to somewhat higher values.

Variation of the speaking rate does not affect all the phonetic units in the same way. The variation of the speaking rate is achieved either by a faster (or slower) movement of the articulators or by omission (or insertion) of some phonetic units. In French, one main adjustment variable of the speaking rate is the pronunciation of the schwa vowel (/ə/). When the articulation rate increases, the length of the schwa can, not only be substantially shortened, but this vowel can completely disappear even in monosyllabic French function words, where the schwa is the only vowel. When the articulation rate is slowing down, there is a tendency to utter all the schwa vowels of a word, and also to add epenthetic schwas after each word final consonant (especially before a pause or a consonantal syllable attack). The schwa vowel that never occurs in a stressed position is generally more affected by duration reduction than the other French vowels; therefore shorter acoustic models are better adapted to their detection.

Figure 3 displays similar statistics for the pronunciation of the final schwa, before a word starting by a consonant, computed on more than 213,000 occur-

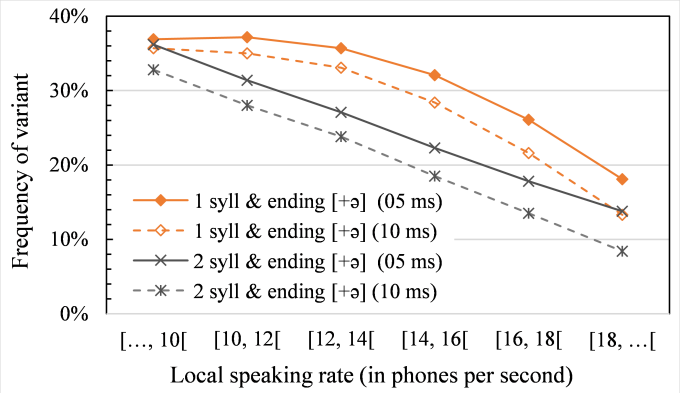

Fig. 3. Frequency of pronunciation of the final schwa, estimated using 5 or 10 ms frame shifts, in one syllable words and in polysyllabic words.

rences of one syllable words such as "*quatre*" (four), "*entre*" (between), and on more than 167,000 occurrences of two or more syllable words such as "*Europe*", "*histoire*" (history), ... Monosyllabic function words, such as "*le*", "*de*", "*que*", ..., are not considered in those statistics. Again, statistics computed using 5 ms frame shifts lead to larger frequency values for the pronunciation of the ending schwa.

Figure 4 concerns the pronunciation of the mute "*e*" in internal positions of words. The statistics displayed here are extracted from 7,500 occurrences of words including two mute "*e*" in internal positions, as for example "*revenir*" (come back) which can be pronounced as:

/ʁ ə v ə n i ʁ/ [+ə] ... [+ə] ...
or /ʁ ə v n i ʁ/ [+ə] ... [-ə] ...
or /ʁ v ə n i ʁ/ [-ə] ... [+ə] ...

According to the pronunciation rules of standard French, when several adjacent syllables contain neutral schwa like vowels, every second schwa can be elided from the pronunciation. This rule is applied especially when the speed of pronunciation is increasing. Also, there is a preference to keep the schwa in the first syllable, in order to avoid consonantal clusters at word attacks and also to place a secondary stress, if possible, on the word first syllable. This preference of maintaining the schwa vowel in first syllables of the words is confirmed by our statistics, where words with elided schwa like vowels in first syllable are very seldom.

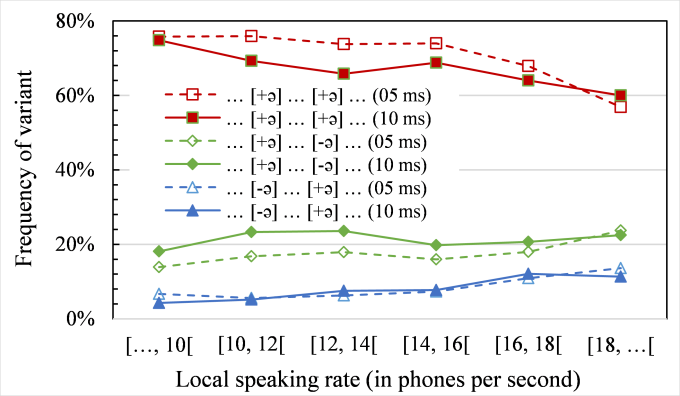

Fig. 4. Frequency of pronunciation of the first and/or the second schwa in polysyllabic words, estimated using 5 or 10 ms frame shifts.

4.2 Analysis of final clusters

The extended lexicon described in Section 3.1 is used here to analyze the speech-text alignments corresponding to clusters /plosive liquid/ in word final position, such as /t ʁ/ at the end of the word "*ministre*" (minister). All possible variants corresponding to the phonetic realization or the elision of any of the phonemes of the cluster are set possible in the extended lexicon, as well as a possible insertion of a final schwa (cf. example for the word "*ministre*" in Table 1). Figure 5 displays the frequency, with respect to the local speaking rate, of some variants observed for the word "*ministre*", which occurs 3,200 times in the data. However, because of a too small number of occurrences associated to the lowest speaking rates ([. . .-10[and [10-12[phones per second), results are reported only for speaking rates higher than 12 phones per second, for which there are more than 400 occurrences in each speaking rate bin.

Results shows that the frequency of the full pronunciation /m i n i s t ʁ ə/ is higher when estimated using the 5 ms frame shift. Two interesting facts are related to the omission of /t/ and of the whole cluster /t ʁ/ in final position at high speaking rates.

The simplification of consonantal clusters especially at word final positions is very frequent in French (but also in other languages). In fact, the final syllabic codas ending with consonants of increasing degree of sonority (example occlusive followed by a liquid) are not favored by the French language and the last liquid can disappear from the pronunciation (example /k a t ʁ ə/ ⇒ /k a t ə/) [6]. The last simplified consonant pronunciation can be strengthened by an additional final schwa like vowel. In more complex word final consonant clusters such as /s t ʁ/ several cluster simplifications can be carried out. As the two adjacent consonants /s/ and /t/ share the same place of articulation, the occlusive can

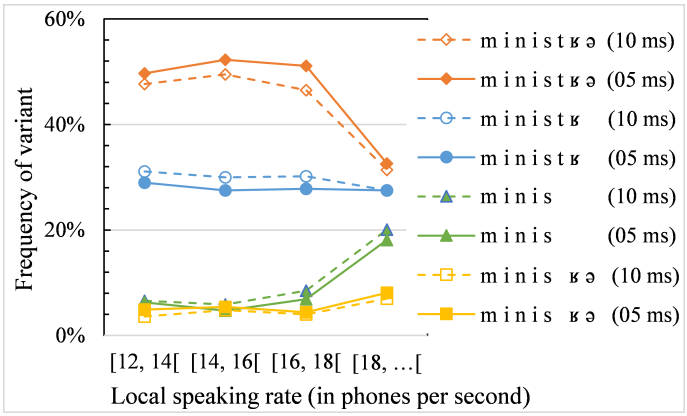

Fig. 5. Frequency of some pronunciation variants for the cluster "*tre*" /t ʁ ə/ at the end of the word "*ministre*" (minister), estimated using 5 or 10 ms frame shifts.

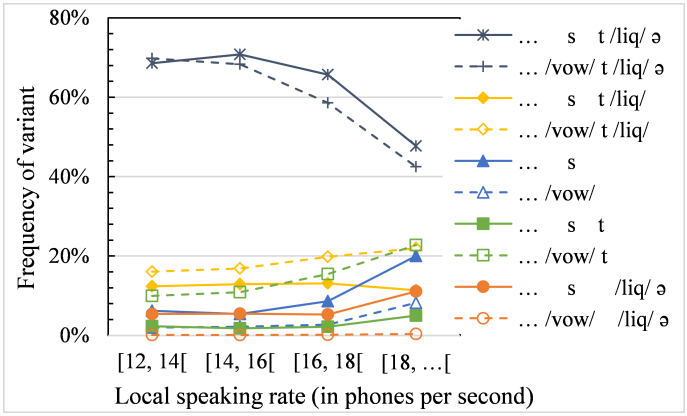

Fig. 6. Frequency of some pronunciation variants for the cluster /t liquid/ in final position of words when this cluster is preceded by a vowel or by the consonant /s/ which has the same place of articulation as /t/.

disappear, with or without the liquid, generating this way several pronunciation variants.

The analysis is then generalized to all the words ending with a cluster of type /t liquid/. To contrast, the analysis statistics are reported separately for the 14,000 occurrences of such words where the cluster is preceded by a vowel sound and for the 2,700 occurrences of words where the cluster is preceded by the consonant /s/ which has the same place of articulation as /t/ (i.e., part of a larger final cluster /s t liquid/).

Figure 6 shows the frequency of omission of the consonant /t/ in the cluster /t liquid/, and the frequency of omission of the whole cluster /t liquid/, are much higher when the preceding sound is the consonant /s/, which has the same place of articulation as /t/, rather than when the preceding sound is a vowel.

5 Conclusion

This paper has investigated the impact of the frame shift used in acoustic analysis on computing corpus-based phonetic statistics through forced speech-text alignment of large speech corpora. The three emitting states of the conventional acoustic hidden Markov models introduce a minimum duration constraint of three frames for each phone segment. With the usual 10 ms frame shift, this corresponds to a 30 ms minimum duration for each phone, which is problematic at high speaking rates. In this paper we compared corpus-based phonetic statistics achieved with a shorter frame shift (5 ms) to those obtained from the usual 10 ms frame shift. Statistics computed on a large speech corpus of more than 3 million running words are analyzed with respect to the various speaking rates. Results exhibit some discrepancies between the two sets of statistics, in particular for high speaking rates, where the usual acoustic analysis frame shift leads to an under-estimation of the frequency of the longest pronunciation variants.

A complementary analysis of pronunciation variants for some word ending clusters was also conducted. The results show that at high speaking rates, final /plosive liquid/ clusters may be omitted, especially when the cluster is preceded by a consonant which shares the same place of articulation as the first consonant (plosive) of the cluster. Further studies will refine such pronunciation statistics, and consider their handling in speech recognition systems.

6 Acknowledgments

This work has been partly realized thanks to the support of the Région Lorraine and the CPER MISN TALC project.

References

1. Adda-Decker, M.: De la reconnaissance automatique de la parole à l´analyse linguistique de corpus oraux. Proc. JEP´2006, XXVIes Journées dEtude sur la Parole, Dinard, France, pp. 389-400 (2006)
2. Adda-Decker, M., and Lamel, L.: Systèmes d´alignement automatique et études de variantes de prononciation. Proc. JEP´2000, XXIIIes Journées d´Etudes sur la Parole 19-23 juin 2000, Aussois, France, pp. 189-192 (2000)
3. Adell, J., Bonafonte, A., Gómez, J. A., and Castro, M. J.: Comparative study of Automatic Phone Segmentation methods for TTS. Proc. ICASSP´2005, IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Philadelphia, USA, pp. 309-312 (2005)

4. Bigi, B., and Hirst, D.: SPeech Phonetization Alignment and Syllabification (SPPAS): a tool for the automatic analysis of speech prosody. Proc. Speech Prosody, Shanghai, China, pp. 1-4 (2012)
5. Bürki, A., Gendrot, C., Gravier, G., Linarès, G., and Fougeron, C.: Alignement automatique et analyse phonétique: comparaison de différents systèmes pour l´analyse du schwa. Traitement Automatique des Langues, 49(3), pp. 165-197 (2008)
6. Côté, M.-H.: Phonetic salience and consonant cluster simplification. B. Bruening, Y. Kang & M. McGinnis, ed. PF: Papers at the interface. MIT Working Papers in Linguistics 29, pp. 229-262 (1997)
7. de Calmès, M., and Pérennou, G.: BDLEX: a Lexicon for Spoken and Written French. Proc. LREC´1998, 1st int. conf. on Language Resources and Evaluation, Grenada, Spain. May 28-30, pp. 1129-1136 (1998)
8. De Mareüil, P. B., Adda-Decker, M., and Gendner, V.: Liaisons in French: a corpus-based study using morpho-syntactic information. Proc. ICPhS´2003, 15th Int. Congress of Phonetic Sciences, Barcelona, Spain (2003)
9. Demuynck, K., and Laureys, T.: A comparison of different approaches to automatic speech segmentation. Proc. Text, Speech and Dialogue, Brno, Czech Republic, pp. 277-284 (2002)
10. EPAC Corpus: Orthographic transcriptions. ELRA catalogue (http://catalog.elra.info), ref. ELRA-S0305
11. Estève, Y., Bazillon, T., Antoine, J.-Y., Béchet, F., and Farinas, J.: The EPAC corpus: Manual and automatic annotations of conversational speech in French broadcast news. Proc. LREC´2010, seventh Int. Conf. on Language Resources and Evaluation, May 19-21, Valetta, Malta (2010)
12. Galliano, S., Gravier, G., and Chaubard, L.: The ESTER 2 evaluation campaign for rich transcription of French broadcasts. Proc. INTERSPEECH´2009, 10th Annual Conf. of the Int. Speech Communication Association, Brighton, UK, September 6-10, pp. 2583-2586 (2009)
13. Goldman, J. P.: EasyAlign: an automatic phonetic alignment tool under Praat. (2011)
14. Gravier, G., Adda, G., Paulsson, N., Carré, M., Giraudel, A., and Galibert, O.: The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. Proc. LREC´2012, 8th Int. Conf. on Language Resources and Evaluation, Istanbul, Turkey, May 23-25 (2012)
15. Huggins-Daines, D., and Rudnicky, A. I.: A Constrained Baum-Welch Algorithm for Improved Phoneme Segmentation and Efcient Training. CMU report (2006)
16. Illina, I., Fohr, D., and Jouvet, D.: Grapheme-to-Phoneme Conversion using Conditional Random Fields. Proc. INTERSPEECH´2011, 12th Annual Conf. of the Int. Speech Communication Association, Florence, Italy, August 27-31 (2011)
17. Jarifi, S., Pastor, D., and Rosec, O.: A fusion approach for automatic speech segmentation of large corpora with application to speech synthesis. Speech Communication, 50(1), pp. 67-80 (2008)
18. Jouvet, D., Fohr, D. and Illina, I.: Detailed pronunciation variant modeling for speech transcription. Proc. INTERSPEECH´2010, 11th Annual Conf. of the Int. Speech Communication Association, Makuhari, Japan, September 26-30 (2010)
19. Kawai, H., and Toda, T.: An evaluation of automatic phone segmentation for concatenative speech synthesis. Proc. ICASSP´2004, IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Montreal, CA, vol. I, pp. 677-680 (2004)
20. Kessens, J. M., and Strik, H.: On automatic phonetic transcription quality: lower word error rates do not guarantee better transcriptions. Computer Speech and Language, 18(2), pp. 123-141 (2004)

21. Kominek, J., and Black, A. W.: A family-of-models approach to HMM-based segmentation for unit selection speech synthesis. Proc. INTERSPEECH´2004, 8th Int. Conf. on Spoken Language Processing, Jeju Island, Korea, October 4-8 (2004)
22. Kuperman, V., Pluymaekers, M., Ernestus, M., and Baayen, H.: Morphological predictability and acoustic duration of interfixes in Dutch compounds. Journal of the Acoustical Society of America, 121(4), pp. 2261-2271 (2007)
23. Mporas, I., Ganchev, T., and Fakotakis, N.: Speech segmentation using regression fusion of boundary predictions. Computer Speech and Language, 24(2), pp. 273-288 (2010)
24. Nakamura, M., Iwano, K., and Furui, S.: Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. Computer Speech and Language, 22(2), 171-184 (2008)
25. Ogbureke, K. U., and Carson-Berndsen, J.: Improving initial boundary estimation for HMM-based automatic phonetic segmentation. Proc. INTERSPEECH´2009, 10th Annual Conf. of the Int. Speech Communication Association, Brighton, UK, September 6-10, pp. 884-887 (2009)
26. Raymond, W. D., Pitt, M. A., Johnson, K., Hume, E., Makashay, M. J., Dautricourt, R., and Hilts, C.: An analysis of transcription consistency in spontaneous speech from the buckeye corpus. Proc. INTERSPEECH´2002, 7th Int. Conf. on Spoken Language Processing, September 16-20, Denver, Colorado, USA (2002)
27. Sjlander, K.: An HMM-based system for automatic segmentation and alignment of speech. Proc. of Fonetik, Lvnger, Sweden, pp. 93-96 (2003)
28. Sphinx. [Online]: `http://cmusphinx.sourceforge.net/` (2011)
29. Toledano, D. T., Gómez, L. A. H., and Grande, L. V.: Automatic phonetic segmentation. IEEE Trans. on Speech and Audio Processing, 11(6), pp. 617-625 (2003)