Speech synthesis in various communicative situations: Impact of pronunciation variations

Sandrine Brognaux [1,2], *Benjamin Picart* [2], *Thomas Drugman* [2]

[1] Cental, ICTEAM - Université catholique de Louvain, Belgium
[2] TCTS Lab - University of Mons, Belgium

sandrine.brognaux@uclouvain.be, benjamin.picart@umons.ac.be, thomas.drugman@umons.ac.be

Abstract

While current research in speech synthesis focuses on the generation of various speaking styles or emotions, very few studies have addressed the possibility of including phonetic variations according to the communicative situation of the target speech (sports commentaries, TV news, etc.). However, significant phonetic variations have been observed, depending on various communicative factors (e.g. spontaneous/read and media broadcast or not). This study analyzes whether these alternative pronunciations contribute to the plausibility of the message and should therefore be considered in synthesis. To this end, subjective tests are performed on synthesized French sports commentaries. They aim at comparing HMM-based speech synthesis with genuine pronunciation and with neutral NLP-produced phonetization. Results show that the integration of the phonetic variations significantly improves the perceived naturalness of the generated speech. They also highlight the relative importance of the various types of variations and show that schwa elisions, in particular, play a crucial role in that respect.

Index Terms: HMM-based speech synthesis, phonetic variations, communicative situation, sports commentaries

1. Introduction

While Text-To-Speech (TTS) synthesis has reached in the last few decades a fairly good level of quality and intelligibility, the neutral carefully read speech it produces is often criticized, as it usually differs from natural human conversations [1, 2, 3]. The need for speech synthesis to better adapt to the context of communication has been recently underlined (e.g. the LISTA project [4]) and the interest has now shifted to the production of speech that suits different speaking styles and emotions. Most studies focussing on this topic have considered modifications at the prosodic and voice quality levels only (cf. [5, 6, 7, 8]). However, surprisingly little attention has been paid to potential phonetic modifications of the sentence to synthesize.

This concern is particularly important in French, where words are characterized by a high amount of phonetic variation (e.g. due to elisions and liaisons phenomena). Many linguistic studies have investigated the factors influencing these variations, such as morpho-syntax [9], speech rate [10, 11], word probability [12], degree of articulation [13], origin of the speaker [14, 15], and so on. The communicative situation (CS), also sometimes referred to as 'phonogenre' [16, 17] (e.g. TV-news speech, political speech) was also shown to play a major role in the phonetic modifications [18].

While modeling phonetic variation has attracted interest in speech recognition (e.g. [19]), few studies have addressed the integration of these variants in speech synthesis. Only the attempt to generate spontaneous-like speech has received attention [1, 2, 3, 20]. Phonetic modeling of some spontaneous speech variation has been proposed in [20, 1] but its impact on speech synthesis has not been evaluated. In [2], a state-based transformation of speech synthesis with Hidden Markov Models (HMM) was presented to convert a neutral synthesized speech signal into spontaneous speech with no modification at the language processing level. Finally, [3] exploited phonetic variation to synthesize German travel information, albeit with a corpus not well suited to spontaneous speech synthesis. They reported, from subjective evaluations, an improvement of the naturalness; however, no consideration was made of the relative importance of the different types of phonetic variation (be they elisions, insertions, or others). Except for this distinction between read and spontaneous speech, phonetic variations have rarely been applied to the synthesis of expressive speech in different CS, with the exception of [21] which proposes to modify the pronunciation of final [ə] only.

This paper builds upon our previous investigations [18] which showed that some communicative 'traits' (e.g. spontaneous vs. read or media-broadcast vs. non-media) are directly correlated with specific phonetic variations like elisions or liaisons. One of the questions this study raised was whether these phonetic variations contributed to the plausibility of the message and should therefore be considered when synthesizing speech with a specific communicative purpose. This paper addresses that concern. To this end, a very large corpus containing about 300 minutes from 32 speakers in 10 CS (e.g. sports commentaries, TV news) is used. First a phonetic analysis of the corpus based on a specific methodology integrating natural language processing (NLP) techniques is presented. Specific attention is paid to live sports commentaries which are further used for synthesis experiments. In the second stage, a subjective evaluation aims at comparing speech synthesized with a neutral read pronunciation and with the real pronunciation of the speaker. The relative importance of the different types of phonetic modifications is also evaluated.

The paper is organized as follows. Section 2 presents the corpus and its annotation. Its analysis in terms of phonetic variations is described in Section 3. The subjective evaluation is presented in Section 4 along with discussion of the results. Finally, Section 5 concludes the paper.

2. Corpus design

Our corpus is an extended version of C-PROM [22] including additional sub-corpora (sports commentaries and corpora designed for TTS purposes). The corpus consists of 315 minutes from 32 French-speaking speakers (French, Belgian and Swiss) and ten sub-corpora covering various CS (interview, political speech, and others). Because they cannot be easily ranked on a single scale, the various CS are defined in this study according to three binary 'situational traits': media/non-media, expressive/non-expressive and read/spontaneous. As these 'traits' are continuums, some corpora were not classified: if their nature regarding a dimension was ambiguous, they were not included in the corresponding set of corpora. The number of speakers per 'trait' ranges from 13 to 17 with an average duration of about 2 hours of speech per trait [1]. In Section 3.3, special attention is paid to two subparts of the corpus : neutral read speech recorded for speech synthesis purposes (READ) and sports commentaries (SPORTS). READ is made of three sub-corpora from different speakers for a total duration of 124 minutes. SPORTS consists of 5 sub-corpora with a total duration of 54 minutes, each corresponding to a different speaker commenting on a basketball, football or rugby match.

For the subjective evaluation (Section 4), an extended version of a basketball commentary from SPORTS ('Sportic' [23]) was used to train an HMM-based speech synthesizer. This corpus is 162 minutes long, silences included. It has the advantage of being spontaneous and of high acoustic quality, and therefore is suited to speech synthesis.

The phonetization of the corpora was done automatically and corrected manually. Corpus and phonetization were then automatically phonetically aligned with EasyAlign [24] and Train&Align [25].

3. Analysis of the phonetic variations

3.1. Methodology

For the phonetic analysis, we developed a specific methodology based on NLP techniques. For each sub-corpus, the orthographic transcription was used to produce its automatic phonetization with the NLP tool eLite [26, 27] designed for TTS purposes. This produced a 'standard' phonetization of the text, corresponding to neutral read speech. This transcription was then automatically aligned with the genuine phonetic transcription as pronounced by the speaker.

This alignment relies on a slightly modified version of Levenshtein's edit distance [18]. All modifications are stored according to their type (insertion, deletion or substitution). Here is an example of the alignment of the sentence "Parce que ça, je pense pas que c'était prévu" with [ə] and liquid deletions :

p	a	R	s	k	@	s	a	Z	@	p	4	s	p	a	k	@	s	e	t	E	p	R	e	v	y
p	a	/	s	k	@	s	a	Z	/	p	4	s	p	a	k	/	s	e	t	E	p	R	e	v	y

The main advantage of using NLP-produced phonetization is that it allows for an easy comparison of the pronunciation of the corpus with a so-called 'standard' pronunciation. The latter already considers most mandatory phonetic variations such as liaisons or elisions dictated by the linguistic context. It provides a more precise analysis than a comparison with a phonetized dictionary while being fully automatic.

The alignment of our corpus with NLP-produced phonetization revealed 4 main types of phonetic modifications which will be further discussed in the next sections:

1. **Schwa elisions**: These are one of the most intricate phonetic variations in French and relate to the optional pronunciation of a [ə] in the middle of a word (e.g. *petite* pronounced [ptit]).
2. **Final schwas**: In French, final schwas may or may not be pronounced at the end of word. We include in this category what we will refer to as 'epenthetic schwas', i.e. the insertion of a final schwa for words not ending in '-e', like the word *match* being pronounced [matʃə] [28].
3. **Final liquid elisions**: This category relates to two phenomena. The first is the elision of the final liquid when preceded by an obstruent (e.g. *peut-être* pronounced [pøtɛt]) [29]. The second is a typical elision of the [l] in the pronoun *il* (meaning *he* or *it*).
4. **Liaisons**: The pronunciation of a latent consonant at the end of a word when followed by a vowel or a mute h (e.g. *les enfants* pronounced [lez̲ɑ̃fɑ̃]).

It should be noted that some of these variations may induce other modifications, in particular on the voicing state of the contextual consonants (i.e. assimilation phenomena). We may, for instance, observe the devoicing of the obstruent when the liquid is elided (e.g. *prendre* pronounced [pʀɑ̃t̠] and not [pʀɑ̃d]) or the voicing of the consonant if the subsequent [ə] is elided (e.g. *ce ballon* pronounced [z̲balɔ̃] instead of [sbalɔ̃]).

3.2. Phonetic variations in different communicative situations

This section presents a brief analysis of the phonetic variation in our corpus according to the three aforementioned situational 'traits'. For this analysis, we follow an approach similar to [18]. Statistical significance of the results is calculated via unilateral t-tests or Wilcoxon tests on the corresponding dataset of each 'trait' (one sample per speaker).

On the whole, the alignment of the genuine phonetics with NLP-produced phonetics shows significantly more variation in spontaneous and non-media speech, compared to read and media speech (with p=1.2e-05 and p=0.043, respectively). With an average of 4.25% modified phonemes (elided, substituted or deleted), spontaneous speech significantly differs from read speech (1.78% modifications). This indicates that, while NLP phonetizers are suited for read speech, they may require some modifications when synthesizing spontaneous or non-media messages. It should be noted that no significant differences are found for the "expressive/non expressive" trait. A possible explanation is that expressive corpora may be characterized by many other factors like the emotion valence (i.e. happy or sad) which could influence their realization.

Regarding the various types of modifications, elisions (both of [ə] and liquids) are significantly more frequent in spontaneous and non-media speech, which confirms findings of [30, 14]. Conversely, liaisons tend to appear more often in read and media speech. These common phonetic tendencies found in spontaneous and non-media speech may be due to their overall higher level of informality compared to read and media speech. A specificity of media speech, however, is the pronunciation of final [ə]. All words are considered here, except for monosyllabic function words (e.g. *le*, *que*, *se*) for which the non-pronunciation of the final [ə] was studied in terms of elision. Although it is somewhat variable across speakers, final [ə] insertion is significantly more frequent (p=0.0028) in media compared to non-media speech, while no significant differences

[1] For a more detailed description of the corpus and of the definition of the situational 'traits' [18].

are observed for the other two situational traits. This confirms the findings of [21] who pointed out a higher rate of ending [ə] pronunciation in radio news and political speech compared to conversational speech. Figure 1 shows the percentage of final [ə] pronunciation in the various CS. It shows that higher rates are observed in media corpora (i.e. sports commentaries, TV news, political speech and interviews).

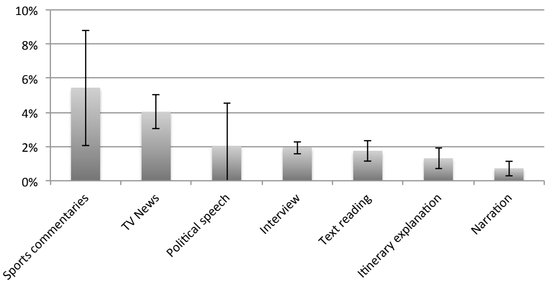

Figure 1: *Percentage of words pronounced with final [ə], for each CS, with 95% confidence intervals.*

3.3. Phonetic variation in sports commentaries

This second stage of our analysis focusses on differences between sports commentaries and neutral read speech recorded for TTS purposes. We expect sports commentaries to display phonetic variation related to spontaneous, media (expressive) speech. Neutral read speech should be, on the other hand, more similar to what is produced by a standard NLP phoneticizer. The aim of our analysis is to highlight the most significant variations observed in sports commentaries to evaluate their impact on speech synthesis in Section 4. The significance of the differences between both corpora is computed on all samples with the tests for equality of two proportions.

Table 1 shows that the percentage of phone variations in sports commentaries is significantly higher than in neutral read speech (p<1e-08). Regarding [ə] elision, our analysis only focusses on non-final [ə] (except for monosyllabic words) as final [ə] may instead be linked to liaisons and is studied separately. As displayed in Table 1, significantly more [ə] elisions are found in sports commentaries (p<1e-08). The analysis of the corresponding occurrences shows that they primarily appear in monosyllabic words; more than 20% of the monosyllabic words ending in -e are shortened (cf. Figure 2). We observe large inter-speaker variations in the elision of [ə] in the middle of a word. Our analysis also showed that mainly function words tend to undergo elision (about 21% against 9% for content words). These two observations are probably correlated as most function words are monosyllabic in French.

Sports commentaries, as expected, also exhibit a significantly higher rate of final [ə] insertion and of epenthetic [ə] (p<1e-08 and p<1e-08, respectively). Further investigations indicated that the insertion of a final [ə] usually occurs before a silence, as studied in [28]. Epenthetic [ə] were observed to appear usually after a consonant, most probably for articulation purposes. The elision of liquids, both in the 'il' pronoun and after an obstruent are also significantly more frequent in sports commentaries (p=1e-08 and p=1.288e-07, respectively), compared to our READ corpus in which they are almost never elided. Finally, as expected in spontaneous speech, sports commentaries display significantly fewer liaisons than TTS neutral read speech (p=5.119e-06). High inter-speaker variability is, however, observed for that measure.

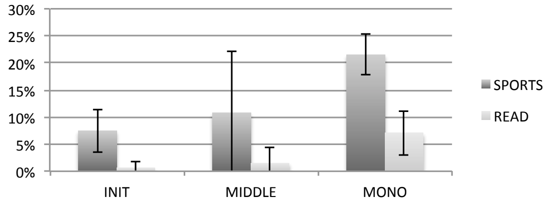

Figure 2: *Percentage of elided [ə] (monosyllabic, initial, or in middle position) in sports commentaries (SPORTS) and neutral read speech for TTS (READ), with 95% confidence intervals.*

4. Subjective evaluation

4.1. Experimental protocol

To evaluate the naturalness of the generated speech with and without specific phonetic variation, an HMM- based speech synthesizer [31] was built, relying on the implementation of the HTS toolkit (version 2.1) publicly available in [32]. For training, 90% of the Sportic corpus [23] was used, leaving 10% for synthesis. The training process relies on manually-corrected phonetic transcriptions. For filter parameterization, we extracted the Mel Generalized Cepstral (MGC) coefficients traditionally used in parametric synthesis. For excitation modeling, the Deterministic plus Stochastic Model (DSM [33]) of the residual signal was used to improve naturalness. Local prosody annotation was used as contextual information in the same way as linguistic information. (For more details about our prosody annotation protocol and its integration in TTS, see [23, 34].)

Two methods of phonetization were compared: use of a standard NLP-produced phonetization (*NLP*), or use of the actually realized pronunciation of the sentence in the corpus (*Real*). This comparison was analyzed through two subjective tests. For the first test, 30 sentences were manually selected from the synthesis set as displaying three or four modifications compared to the NLP-produced phonetization. Around 20% of the sentences of the corpus contain 3 or more modifications. Occasional errors made by the NLP phoneticizer were manually corrected in order not to influence the listener. 22 native French-speaking participants, mainly naive, took the test. For 15 randomly selected sentences, two versions were compared: the real and the NLP phonetization. Listeners were asked to score the naturalness of the message as corresponding to sports commentaries. The scale ranged from -3 (much less natural) to +3 (much more natural). A score of 0 was to be chosen if both versions were seen as equivalent.

The purpose of the second test was to assess the relative role played by 4 types of phonetic variation: [ə] elision, final [ə] insertion, liquid elision and liaison. 60 sentences (or chunks of sentences) were manually selected from the synthesis test, i.e., 15 for each type of phonetic variation. Each sentence originally contained only one phonetic modification of the corresponding type. Another 21 native French participants, mainly naive, took the test. They were asked to compare 16 randomly selected pairs of sentences, 4 for each type of phonetic variation. Other evaluation conditions were the same as in the first test. For the interpretation of the results, statistical significance is computed with Wilcoxon signed rank tests comparing the average percent-

Table 1: *Phonetic changes for sports commentaries (SPORTS) and neutral read speech for TTS (READ).*

	All changes	[ə] elision	Final [ə] pronunciation	Epenthetic [ə]	[l] elision in ‘il’	Elision of liquid after obstruent	Liaison
READ	1.34%	17.33%	2.20%	0.13%	0%	5.68%	55.90%
SPORTS	2.94%	5.92%	5.43%	2.98%	41.48%	12.34%	47.21%

age of preferences of the 22 (or 21) testers.

4.2. Evaluation and discussion

Results of the first test are displayed in Figure 3, which shows a significant degree of preference for the genuine phonetization containing phonetic variation (*p*=0.012). Synthesis with real phonetization is preferred in 45.15% of the cases against 31.21% for the NLP-produced phonetization, while the two were found to be equivalent in 24% of cases (average CMOS score of 0.26). Figure 4 highlights the significantly higher percentage of preference for the real pronunciation in sentences in which at least 2 elisions appear (*p*=1.86e-04). In this case, real phonetics are preferred in 66.67% of the cases against 19.44% for NLP-produced phonetization (average CMOS score of 0.81). Conversely, sentences with more than 2 final schwa insertions were assigned lower scores. This indicates that [ə] elisions may be required, as opposed to final [ə] insertion. This is investigated further in our second subjective test.

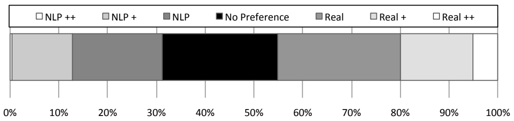

Figure 3: *Preferences (%) for NLP-produced and real phonetization for the first test. “+” refers to stronger preference scores.*

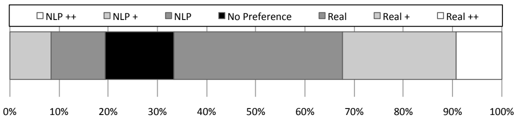

Figure 4: *Preferences (%) for NLP-produced and real phonetization for the first test in sentences with at least two [ə] elisions.*

Overall results of the second test show substantial differences in ratings according to the type of phonetic variation (cf. Figure 5). It should first be pointed out that the overall preference for the real phonetization is slightly lower than in the first experiment, which may be explained by the fact that both versions of the test sentences only differ by a single variation. This may imply that the use of such short chunks of speech, out of context and independent of other variation, somehow influences the results and thus may not reflect a real situation. This should prevent us from drawing hasty conclusions.

For liaisons and final [ə], the NLP-produced phonetization tends to be preferred, i.e. a realized liaison and no insertion of final [ə]. For final [ə] insertion, the real phonetization is however slightly preferred in the case of epenthetic [ə], i.e. insertion of final [ə] in words not ending in -e. This phenomenon was indeed shown to be highly specific to sports commentaries [18]. A more detailed analysis of the prosody patterns of such insertions might allow for a better realization of that type of variation. For the elision of liquids, the real phonetization tends to be preferred, i.e. with elision of the final liquid after an obstruent or in the ‘il’ pronoun. Further investigation shows that this rate is preferred (46.67% vs. 33.33% for NLP-produced phonetization) when considering only final sequences with obstruent and liquid; the variation in the ‘il’ pronoun usually results in a ‘no preference’ rate. This difference, however, is not statically significant (*p*=0.06). Finally, as noticed in the first experiments, the real phonetization is rated significantly more natural than the NLP-produced pronunciation for the elision of [ə] (*p*= 0.006). Further studies should also investigate the role played by these variations on intelligibility.

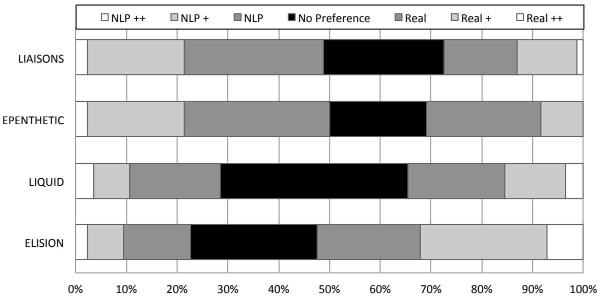

Figure 5: *Preferences (%) between NLP-produced and real phonetization according to the type of variation*

5. Conclusion

The goal of this study was to assess whether the phonetic variation that appears in various communicative situations (CS) contributes to the plausibility of the message and should, in that respect, be integrated in speech synthesis. Our phonetic analysis of a large French corpus with various CS showed that substantial phonetic variation is observed according to the communicative context. Sports commentaries, in particular, are characterized by higher rates of [ə] and liquid elisions, more final [ə] pronunciations and fewer liaisons. Subjective evaluations with HMM-based synthesized extracts of sports commentaries indicate that, on the whole, integrating phonetic variants significantly improves the naturalness of the synthesized message. While the insertion of final [ə] or the reduction of the number of liaisons does not enhance the naturalness, integrating [ə] elision and, to a lesser extent, final liquid elision after obstruents turns out to improve the overall naturalness. This indicates that the synthesis of French sports commentaries should benefit from such phonetic modification. Further studies should be done to corroborate whether or not these conclusions also hold for other speaking styles and languages.

6. Acknowledgement

S. Brognaux and T. Drugman are supported by FNRS. The project is partly funded by Walloon Region Wist 3 SPORTIC.

7. References

[1] K. Prahallad, A. Black, and R. Mosur, “Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis,” in *ICASSP*, 2006.

[2] C.-H. Lee, C.-H. Wu, and J.-C. Guo, “Pronunciation variation generation for spontaneous speech synthesis using state-based voice transformation,” in *ICASSP*, 2010.

[3] S. Werner and R. Hoffmann, “Pronunciation variant selection for spontaneous speech synthesis - a summary of experimental results,” in *Speech Prosody*, 2006.

[4] M. Cooke, S. King, M. Garnier, and V. Aubanel, “The listening talker: A review of human and algorithmic context-induced modifications of speech,” *Computer Speech and Language*, vol. 28(2), pp. 543–571, 2014.

[5] K. Miyanaga, T. Masuko, and T. Kobayashi, “A style control technique for HMM-based speech synthesis,” in *ICSLP*, 2004.

[6] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, “Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis,” *IEICE Transactions on Information and Systems*, vol. 88, no. 3, pp. 502–509, 2005.

[7] R. Tsuzuki, H. Zen, K. Tokuda, T. Kitamura, M. Bulut, and S. Narayanan, “Constructing emotional speech synthesizers with limited speech database,” in *ICSLP*, 2004.

[8] N. Obin, P. Lanchantin, A. Lacheret, and X. Rodet, “Discrete/continuous modelling of speaking style in HMM-based speech synthesis: Design and evaluation,” in *Interspeech*, 2011.

[9] P. Boula de Mareuil, M. Adda-Decker, and V. Gendner, “Liaisons in French: a corpus-based study using morpho-syntactic information,” in *ICPhS*, 2003.

[10] A. Lacheret-Dujour, “Phonological variations in read speech, reduction phenomena and speaker classes: do allophonic choices represent speaking style?” in *ESCA Workshop on Phonetics and Phonology of Speaking Styles*, 1991.

[11] C. Fougeron, J.-P. Goldman, A. Dart, L. Gulat, and C. Jeager, “Influence de facteurs stylistiques, syntaxiques et lexicaux sur la réalisation de la liaison en français,” in *Actes de TALN*, 2001.

[12] D. Jurafsky, A. Bell, M. Gregory, and W. D. Raymond, “Probabilistic relations between words: Evidence from reduction in lexical production,” *Typological studies in language*, vol. 45, pp. 229–254, 2001.

[13] B. Picart, T. Drugman, and T. Dutoit, “Analysis and synthesis of hypo and hyperarticulated speech,” in *7th Speech Synthesis Workshop (SSW7)*, 2010.

[14] F. Hambye, “La prononciation du français contemporain en Belgique. variations, normes et identités.” Ph.D. dissertation, Université catholique de Louvain, Belgique, 2005.

[15] A. Martinet, *La prononciation du français contemporain*. Librairie Droz, 1971.

[16] J. P. Goldman, A. Auchlin, and A. C. Simon, “Discrimination de styles de parole par analyse prosodique semi-automatique,” in *Interface Discours Prosodie (IDP)*, 2009.

[17] T. Pršir, J. P. Goldman, and A. Auchlin, “Variation prosodique situationnelle: étude sur corpus de huit phonogenres en français,” in *Interface Discours Prosodie (IDP)*, 2013.

[18] S. Brognaux and T. Drugman, “Phonetic variations: Impact of the communicative situation,” in *Speech Prosody*, 2014.

[19] H. Strik and C. Cucchiarini, “Modeling pronunciation variation for asr: A survey of the literature,” *Speech Communication*, vol. 29, no. 2, pp. 225–246, 1999.

[20] C. Bennett and A. Black, “Prediction of pronunciation variations for speech synthesis: A data-driven approach,” in *ICASSP*, 2005.

[21] S. Roekhaut, J.-P. Goldman, and A. C. Simon, “A model for varying speaking style in TTS systems,” in *Speech Prosody*, 2010.

[22] M. Avanzi, A. C. Simon, J. P. Goldman, and A. Auchlin, “C-PROM. An annotated corpus for French prominence studies.” in *Speech Prosody*, 2010.

[23] S. Brognaux, B. Picart, and T. Drugman, “A new prosody annotation protocol for live sports commentaries,” in *Interspeech*, 2013.

[24] J.-P. Goldman, “EasyAlign: An automatic phonetic alignment tool under Praat,” in *Interspeech*, 2011.

[25] S. Brognaux, S. Roekhaut, T. Drugman, and R. Beaufort, “Train&Align: A new online tool for automatic phonetic alignments,” in *IEEE Workshop on Spoken Language Technologies (SLT)*, 2012. [Online]. Available: http://cental.fltr.ucl.ac.be/train_and_align/

[26] V. Colotte and R. Beaufort, “Linguistic features weighting for a text-to-speech system without prosody model,” in *Interspeech*, 2005.

[27] R. Beaufort and A. Ruelle, “eLite : système de synthèse de la parole à orientation linguistiques,” in *Journées d’étude sur la parole (JEP)*, 2006.

[28] A. Hansen, “The covariation of [schwa] with style in Parisian French: an empirical study of ‘E caduc’ and prepausal [schwa],” in *ESCA Workshop on Phonetics and Phonology of Speaking Styles*, 1991.

[29] J. W. de Reuse, “La phonologie du français de la région de Charleroi (Belgique) et ses rapports avec le wallon,” *La linguistique*, vol. 23, pp. 99–115, 1987.

[30] C. Fougeron, J.-P. Goldman, and U. H. Frauenfelder, “Liaison and schwa deletion in French: an effect of lexical frequency and competition?” in *Interspeech*, 2001.

[31] H. Zen, K. Tokuda, and A. Black, “Statistical parametric speech synthesis,” *Speech Communication*, vol. 51(11), pp. 1039–1064, 2009.

[32] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, “The HMM-based speech synthesis system (HTS) version 2.0,” in *Sixth ISCA Workshop on Speech Synthesis (SSW6)*, 2007.

[33] T. Drugman and T. Dutoit, “The deterministic plus stochastic model of the residual signal and its applications,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20(3), pp. 968–981, 2012.

[34] B. Picart, S. Brognaux, and T. Drugman, “HMM-based speech synthesis of live sports commentaries: Integration of a two-layer prosody annotation.” in *8th ISCA Speech Synthesis Workshop (SSW8)*, 2013.