A New Prosody Annotation Protocol for Live Sports Commentaries

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

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

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

Abstract

This paper proposes a new prosody annotation protocol specific to live sports commentaries. Two levels of annotation are defined with HMM-based speech synthesis in view. Local labels are assigned to all syllables and refer to accentual phenomena. Global labels classify sequences of words into five distinct sub-genres, defined in terms of valence and arousal. The objective of the study is to provide a set of labels both related to a specific function and characterized by a distinct acoustic realization. The consideration of these constraints should allow for an automatic prediction of the labels both from the text or from the speech signal. Reasonable inter-annotator scores are achieved for both annotation levels. A prosodic analysis of all labels also shows that they can usually be distinguished by specific acoustic realizations. The integration of this new annotation protocol within HMM-based speech synthesis shows promising results.
Index Terms: Prosody, Expressive speech synthesis, Sports

1. Introduction

Speech synthesis forms part of our everyday life and has a wide number of practical applications (GPS, games, electronic cards, etc.). However, the discrepancies that still remain between a synthetic voice and a natural human voice preclude its large-scale marketing. Its lack of naturalness is notably triggered by the inability of existing systems to deal with expressivity, which is either not modeled at all or only in a very limited way [1]. The generation of an expressive prosodic realization is of utmost importance when synthesizing sports commentaries. Their high degree of expressivity has been shown to fulfill various functions (e.g. expressing excitation, frustration, catching attention, etc.) [2][3] which should be reproduced in the synthetic voice.

Several studies have focused on the prosodic analysis of sports commentaries (i.e. basketball, football and rugby [4], horse races [5], soccer [6] and football [2] [3]). This phonostyle has been shown to display a very specific prosodic realization that highly differs from other styles [6]. It is characterized by expressive accentual patterns and most studies also emphasize its division into several speech styles (or speaking styles) [3] (Figure 1). These speech styles will be referred to as 'sub-genres' in the remainder of the paper and relate to what others have called 'prosodies' [7] or 'discourse ambiances' [8]. Whilst the *elaboration* corresponds to a somehow neutral speaking style, dramatic speech is characterized by a high arousal degree, which rises during the *building up suspense* and reaches a climax at the *presentation of a highlight*. These sub-genres are characterized by a specific function and by a rather stable acoustic realization. [4] shows, for example, that highlight events like shots are usually realized with a significantly higher fundamental frequency (F0). The analysis of horse race commentaries [5] shows similar results for high emotional phases like the end of the race. The analysis of sequences happening just after a goal in football commentaries further indicates that the prosodic realization depends whether the goal is for or against the supported team [2]. Sports commentaries being highly 'listener-oriented', these two distinct acoustic realizations help the listener decode the message more quickly. Conversely to the classification of [3], the distinction is based on valence rather than on the arousal level. On the whole, most studies tend to suggest that a prosody annotation of sports commentaries requires, beside local accentual information, a more global annotation level assigning a specific sub-genre to the speech segments.

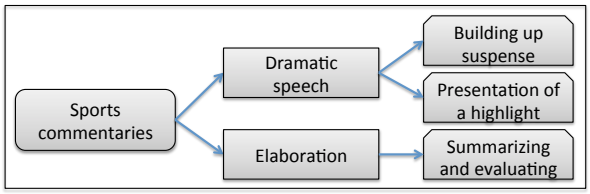

Figure 1: *Speech styles in sports commentaries [3]*

Various accentuation models have been developed with speech synthesis in view and could be exploited to define a local annotation of expressive speech. ToBI [9] proposes a detailed prosody annotation system. It should however be highlighted that the complexity of the system makes it difficult to predict the labels from the text. Mertens [10] also presented a model specific to French, aiming at discretizing the prosodic continuum into pitch levels assigned to the syllables. It was, however, originally designed for neutral speech and presents some drawbacks when exploited for expressive speech, notably for non-initial emphatic stresses. The global annotation in sub-genres could, on the other hand, be based on [3] and [2].

The objective of this paper is to present a prosody annotation protocol specific to sports commentaries (basketball in particular) and relying on two annotation levels. A local annotation is associated to the syllable level and aims at annotating accentual events. A global annotation classifies groups of words into specific sub-genres. This annotation protocol was developed with HMM-based speech synthesis in view. The local labels are included in the contextual information provided to the system whilst the global annotation is exploited to train distinct models for each sub-genre. The use of this annotation for speech synthesis implies several constraints. The local and global labels should be associated with a specific expressive function. Assuming that a semantic analysis of the text is available, it would then be easy to predict the labels from the text. Secondly, to avoid averaging very different acoustic realizations, each label should be characterized by specific acoustic features.

The paper is organized as follows. Section 2 presents the

corpus used throughout this study. Section 3 further details the proposed annotation protocol. Section 4 provides an acoustic analysis of both annotation levels (global and local). The integration of the proposed annotation protocol within a HMM-based speech synthesis system is investigated in Section 5. Finally, Section 6 concludes and presents further developments.

2. Corpus of sports commentaries

This study is based on a corpus of live commentaries of two basketball games, uttered by a professional French commentator and recorded in sound-proof conditions. The speaker watched the game and commented it without any prompt. The issue with sports commentaries corpora is usually the high level of background noise which precludes their precise acoustic analysis [2]. Conversely, our corpus exhibits the advantage of being spontaneous and of high acoustic quality, being therefore suited for speech synthesis. Both matches star the Spirou Belgian team with very tight final scores, which induces a high level of excitation. The corpus lasts 162 minutes, silences included.

The corpus was orthographically and phonetically transcribed by [11] with manual check. The phonetic transcription was aligned with the sound with [12], exploiting the bootstrapping function to achieve alignment rates around 85% with a 20 ms threshold. Finally, other linguistic information (syllables, parts of speech, rhythmic groups, etc.) were generated by [13].

3. Proposed annotation protocol

3.1. Prosody annotation protocol

A two-tier prosody annotation is proposed. A local tier, linked to the syllable level, refers to accentual phenomena. It contains a small amount of labels that can be predicted from the text (Table 1). Each label fulfills a distinct and specific function. Five labels are related to non-emphatic stresses [14] and are assigned to the end of accentual phrases. They are characterized by a pitch level, H for rising or higher pitch vs. L for falling or lower pitch. They can also be distinguished by the level of boundary they determine, similarly to boundary tones in [9]. To facilitate the automatic annotation of the labels (from the text or from the acoustics), these two levels are distinguished according to the presence or absence of a subsequent silence. Conversely to H and L syllables, HH and LL syllables are directly followed by a silent pause. A specific tag E is assigned to the final boundary of player names enumerations, which are very common in sports commentaries and may display a specific acoustic realization. A focus stress (F) relates to emphatic stresses. An hesitation label (He), and a creaky label (C) allow avoiding the degradation of the models at training time. Indeed, hesitations are realized with long durations whilst creaky syllables are characterized by a very low pitch [15]. If these syllables are not singled out, their prosodic features may influence the synthesized prosody. All remaining syllables are assigned a NA symbol.

Table 1: *List of local labels*

Stresses			
Not emphatic	Emphatic	Unstressed	Other
H HH L LL E	F	NA	He C

The global tier, inspired by [3] and [2], is assigned to groups of words, conversely to the local labels which are assigned to each syllable. It basically classifies the speech segments into sub-genres, based on a dimensional analysis of emotions [16]. The valence and the arousal levels drove us to define five sub-genres (Figure 2). While Excited and ExMax correspond to positive excitation, NegTension is linked to frustration, e.g. for failed shots. ExRise relates to a dynamic rising of the excitation.

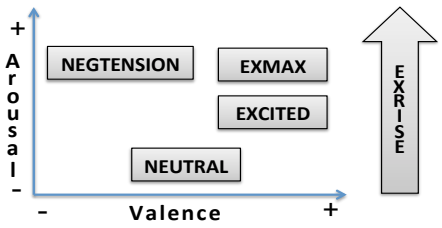

Figure 2: *The global labels on a dimensional scale*

The corpus was manually annotated with the two-tier annotation. Only one stress-related label could be assigned to each syllable, possibly associated with He or C. If an emphatic stress fell on an accentual phrase boundary, the F label overwrote the non-emphatic stress. The annotation results from two or three listenings of each sequence of 4-5 words and was submitted to a second check by the same annotator. This was performed with Praat [17] with acoustic cues displayed (F0, intensity and formants). The annotator was required to rely on his auditory senses only, except if the acoustic cues completely contradicted his perception. For local labels, it allowed avoiding the annotation of stresses with no specific acoustic realization, which would degrade the quality of the synthesis. Indeed, it has been shown [18] that human annotators tend to overdetect prominences in specific grammatical locations where a stress is expected, or underdetect them on clitic words like determiners. We therefore carefully tried to avoid such issues, the objective of our annotation protocol being to assess the acoustic realization and not to offer a perceptive representation of the corpus.

3.2. Inter-annotator rate

Twenty percents of the corpus were annotated by a second expert, the initial annotation being hidden. The local tier reaches a Cohen's kappa score [19] of 0.66, with an observed agreement of 80.22%. This is comparable to the inter-annotator rate achieved by ToBI [9]. For the global tier, a lower kappa score was achieved, i.e. 0.38, with an observed agreement of 54.33%. As for the local annotation, this score was computed at the syllable level. It should however be noted that the confusion matrix shows logical confusions between the sub-genres (see Table 2). Unsurprisingly, *Neutral* and *Excited* (as well as *ExMax* and *Excited*) tend to be interchanged, the distinction between those three labels proposing a discretization of the arousal continuum.

Table 2: *Inter-annotator confusion matrix for the global annotation of the corpus in sub-genres (NegT = NegTension)*

	Neutral	Excited	ExMax	NegT	ExRise
Neutral	1047	147	0	77	17
Excited	869	499	144	53	28
ExMax	14	20	152	24	0
NegT	122	68	17	247	3
ExRise	97	124	58	9	305

4. Acoustic analysis of the prosodic labels

This section shows to what extent our labels are characterized by distinct acoustic realizations, which could then be reflected in the synthetic speech.

4.1. Analysis of the local annotation

Table 3 displays the average acoustic values for four discriminant prosodic features. Duration measures are extracted with [20]. The total perceptual loudness [21] and the fundamental frequency (using the SRH [22] algorithm) are also analyzed. Dynamic values are computed as the difference between the value of the last and first frames of 10 ms of the syllable.

Table 3: *Average acoustic values of the local labels*

	Nucleus Dur (sec)	Mean F0 (Hz)	Dyn F0 (Hz)	Dyn Loudness (dB)
H	0.07	223	10.1	-2.2
HH	0.14	222	9.2	-26.2
E	0.11	184	19.5	-16.5
L	0.10	221	-18.1	-2.7
LL	0.12	202	-42.3	-26.0
F	0.07	245	44.5	5.8
He	0.18	185	-11.6	-3.3
C	0.03	170	-6.5	13.6
NA	0.05	227	-2.7	9.2

It should first be noted that the nucleus duration helps distinguishing between intonational boundaries followed (LL and HH) and not followed (L and H) by a silence. The duration of focused nuclei (F) remains fairly low. On the other hand, the mean pitch for F is higher than for all other syllables. This confirms the widespread assumption stating that emphatic stresses are more realized in terms of pitch than lengthening, conversely to boundary stresses [23]. The dynamic F0 distinguishes between rising/high (or continuative) and falling/low (or conclusive) non-emphatic stresses. Emphatic stresses in our corpus show a very much rising pitch. Finally, the loudness dynamics draw the distinction between emphatic and non-emphatic stresses, the latter displaying a falling loudness pattern because of their location at the end of an accentual phrase. Like HH, E is characterized by a rising pitch, nucleus lengthening and falling loudness. It is, however, realized with a clearly lower pitch.

A detailed study of the emphatic stresses in our corpus highlighted several interesting points. Conversely to non-emphatic stresses, emphatic stresses tend to fall on numbers (in 27% of the cases against 9% for non-emphatic stresses). This finding is clearly specific to sports commentaries. Numbers often refer to scores and play a significant role in the expressive function. It was also found that emphatic stresses tend to fall on the first syllable of the word (as stated by [24]) conversely to non-emphatic stresses. They are also often preceded by a silence.

A last side of our study focused on the comparison between the local labels and annotations provided by automatic prominence detection tools. Prosoprom [25] automatically annotates each syllable as prominent or not. The annotation relies on rules learned on a French annotated corpus and considers various prosodic features (F0, duration, silences, etc.). 42.1% of our stressed syllables (i.e. H, HH, L, LL, E and F) are regarded as prominent by Prosoprom, against 10.9% for the unstressed syllables. F-labelled syllables achieve the highest percentage (60.6%). Prosoprom was recently developed into a gradual prominence annotation tool [26] (PromGrad), which defines five prominence levels, ranging from 0 to 4. Table 4 shows the average prominence value assigned to each label of our corpus. The two boundary levels related to non-emphatic stresses are clearly reflected by the automatic annotation. It should be noted that PromGrad defines a syllable as prominent based on a complex value combining the various prosodic parameters (i.e. mean pitch, duration, pitch movement and duration of the following pause). This means that, if a stress is realized by few distinct features, it is less likely to be assigned a high prominence level. It is here clearly the case for emphatic stresses that are less characterized by pitch and are usually not followed by a pause. An algorithm considering the presence of a preceding pause might alleviate this issue. It should also be highlighted that this clear relation between our annotation and [26] implies that it could be possible, to a certain extent, to automatically predict the local labels from the acoustic realization.

Table 4: *Average prominence levels assigned by [26]*

H	HH	E	L	LL	F	He	C	NA
0.9	3.2	2.7	0.7	2.8	1.5	1.4	0.3	0.2

4.2. Analysis of the global annotation

An analysis of the prosodic realization of each global style is shown in Table 5. The ratio of rises corresponds to the percentage of rising syllables. The articulation rate is computed as the number of syllables divided by the articulation duration. Finally, the ratio of silences is computed as the total duration of silences in each sub-genre divided by the total duration of this sub-genre. It should be noted that initial and final silences are not considered as they could be arbitrarily assigned to the previous or following sub-genre.

Table 5: *Average acoustic values of the global tags*

	Mean F0 (Hz)	Mean Loudness (dB)	Ratio Rise (%)	Articulation Rate	Ratio Sil (%)
Neutral	212	29.7	2.6	5.6	46
Excited	227	33.4	2.3	6	32.8
ExMax	261	41.1	1.5	5.2	18
NegTension	215	32.6	2.7	5.6	32.7
ExRise	215	29.4	15.5	5.5	29.5

A first insightful observation concerns the higher pitch value of the *ExMax* sub-genre. This finding is in line with other studies pointing out higher F0 values during and after shots [4] [2], which are, in our corpus, often assigned an *ExMax* label. As expected, *Excited* is realized with a pitch frequency lying between *Neutral* and *ExMax*. Similar observations can be made for the loudness, with higher values for *ExMax* and intermediate values for *Excited*. Noteworthy, however, is that *NegTension* is characterized by a relatively low pitch, despite its high arousal level on the dimensional scale. This confirms the assumption of [2] which states that disappointment should lead to a depressed pitch value. The ratio rise clearly draws a distinction between *ExRise* and the other sub-genres, *ExRise* displaying a continuous rising of the pitch. Conversely to what might be expected, the articulation rate is rather stable across all styles. This phenomenon has already been noticed in other studies on sports

commentaries [5] [3]. However, we observe a clear reduction of the ratio of silences, according to the arousal level, which indicates a rise of the speaking rate, when including silences.

We also analyzed the probability of transition from one sub-genre to another. As expected, *ExRise* usually leads to *ExMax* (0.38) or *NegTension* (0.38), depending on the success or failure of the action and on the implicated team. The end of *ExMax* and *NegTension* usually indicates the end of the action and is followed by a *Neutral* tag. This transition matrix, along with the average duration of each sub-genre segment, may be considered when automatically selecting global tags at synthesis time.

Finally, the evolution of the acoustic features was studied. The objective was to investigate whether a clear change was noticed at the intersection between two sub-genres, as this would facilitate an automatic annotation of new corpora based on the acoustics only. Prosodyn [8] produces such a representation of the prosodic dynamism of a speech segment. The average value of the prosodic features is computed on a sliding window of 19 syllables. Its application on a representative speech file of the corpus is shown in Figure 3, with the segmentation in sub-genres at the bottom. The evolution of the pitch, in semi tones, is shown to clearly follow the arousal level.

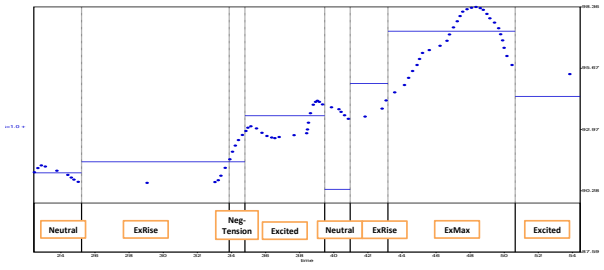

Figure 3: *Dynamic representation of the pitch on a small part of the corpus with Prosodyn [8]*

5. Application to speech synthesis

In order to assess the validity of our local and global labels definition, several HMM-based speech synthesizers [27] were built, relying on the implementation of the HTS toolkit (version 2.1) publicly available in [28]. The first synthesizer (*Syn1*) is the baseline system, in which a manually-checked phonetic transcription is the only contextual information. The second synthesizer (*Syn2*) makes use of the same phonetic transcription, but complemented with the specific information at the local level. The third (*Syn3*) and fourth (*Syn4*) synthesizers are similar respectively to the first and second systems, but differ by the fact that they consist of several sub-synthesizers (one per global label) which are trained on exclusive subsets of the whole corpus specific to the global label they correspond to (i.e. to the sub-genre they model). For each synthesizer, 90% of the corresponding database was used for the training, leaving around 10% for the synthesis. As filter parameterization, we extracted the Mel Generalized Cepstral (MGC [29]) coefficients traditionally used in parametric synthesis. As excitation modeling, the Deterministic plus Stochastic Model (DSM [30]) of the residual signal was used to improve naturalness.

The quality of the resulting synthesis was perceptively assessed through informal subjective tests. The integration of the local tier (*Syn2*) achieves high-quality results, with a clear improvement in the realization of stresses. Focusses and hesitations are particularly well generated. Figure 4 shows that the integration of accentual information allows producing, for a focussed syllable, a higher and more dynamic pitch contour, along with a lengthening of the syllable. The integration of global information (*Syn3* and *Syn4*) produces, as expected, distinct models with specific global prosodic features. The *ExMax* model is characterized by higher pitch and loudness levels, whilst the *NegTension* model presents a clearly lower average pitch. The signal, is however of poorer segmental quality than for the two first synthesizers. This can be explained by the smaller amount of data available to train the models. The use of adaptation techniques to train independent models of each sub-genre, based on the *Neutral* model or on an average model trained on all sub-genres, should alleviate this problem. The integration of both annotation layers already shows promising results.

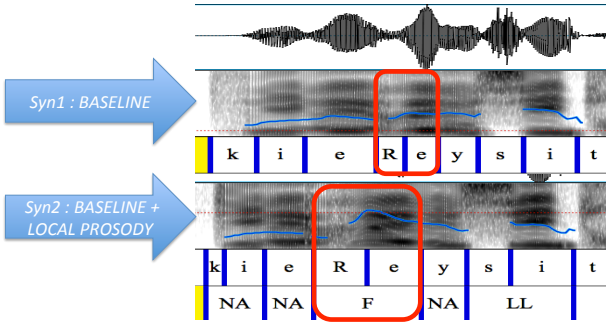

Figure 4: *Synthesis (Syn1 and Syn2) of a sentence with focus*

6. Conclusion and perspectives

This paper proposed a prosody annotation protocol designed for sports commentaries with HMM-based speech synthesis in view. Two annotation levels are defined to represent both the accentual patterns and the various sub-genres specific to sports commentaries. The advantage is that the labels are characterized by a distinct expressive function and by a rather stable prosodic realization. This was shown on a 2 hour-long corpus of basketball commentaries and should allow the automatic prediction of the labels both from the text or from the speech signal. The local labels are included in the contextual information given to the synthesizer whilst the global labels allow for the training of distinct models for each sub-genre. A first perceptive analysis shows that a clear improvement is achieved when integrating the local prosodic information. The consideration of the sub-genre annotation produces models with specific prosodic features which generate a more appropriate global prosody.

Perspectives include the use of adaptation techniques to train the models of each sub-genre and the automatic prediction of the labels from text only (at synthesis time) and from text and acoustics (at training time). Further research will investigate the application of the annotation protocol to other sports corpora.

7. Ackowledgements

Authors are supported by FNRS. The project is partly funded by the Walloon Region Wist 3 SPORTIC. Authors are grateful to S. Audrit for her implication in the recording of the corpus, and to A. C. Simon and M. Avanzi for their insightful advice.

8. References

[1] N. Campbell, "Conversational speech synthesis and the need for some laughter," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 14(4), pp. 1171–1179, 2006.

[2] J. Trouvain, "Between excitement and triumph - live football commentaries in radio vs. tv," in *17th International Congress of Phonetic Sciences (ICPhS XVII)*, 2011.

[3] F. Kern, *Prosody in Interaction*. John Benjamins, 2010, ch. Speaking Dramatically. The Prosody of Live Radio Commentary of Football Matches, pp. 217–237.

[4] S. Audrit, T. Psir, A. Auchlin, and J.-P. Goldman, "Sport in the media: A contrasted study of three sport live media reports with semi-automatic tools," in *Speech Prosody*, 2012.

[5] J. Trouvain and W. Barry, "The prosody of excitement in horse race commentaries," in *ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, 2000, pp. 86–91.

[6] N. Obin, V. Dellwo, A. Lacheret, and X. Rodet, "Expectations for discourse genre identification," in *Interspeech*, 2010.

[7] R. Odgen, "We speak prosodies and we listen to them," in *Symposium on Prosody and Interaction*, 2001.

[8] J.-P. Goldman, "Prosodyn: a graphical representation of macro-prosody for phonostylistic ambiance change detection," in *Speech Prosody*, 2012.

[9] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "Tobi: A standard for labeling english prosody," in *International Conference on Spoken Language Processing (ICSLP)*, 1992, pp. 867–870.

[10] P. Mertens, "L'intonation du français. de la description linguistique la reconnaissance automatique." Ph.D. dissertation, Univ. Leuven (Belgium), 1987.

[11] J.-P. Goldman, "Easyalign: an automatic phonetic alignment tool under Praat," in *Interspeech*, 2011, pp. 3233–3236.

[12] 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*, 2012.

[13] V. Colotte and R. Beaufort, "Linguistic features weighting for a text-to-speech system without prosody model," in *Interspeech*, 2005, pp. 2549–2552.

[14] A. Di Cristo, "Vers une modélisation de l'accentuation du français : deuxième partie," *Journal of French Studies*, vol. 10, pp. 27–44, 2000.

[15] T. Drugman, J. Kane, and C. Goble, "Resonator-based creaky voice detection," in *Interspeech*, 2012.

[16] A. Mehrabian and J. A. Russel, *An Approach to Environmental Psychology*. MIT Press, 1974.

[17] P. Boersma and D. Weenink. (2009, May) Praat: doing phonetics by computer (version 5.1.05) [computer program]. [Online]. Available: http://www.praat.org

[18] J.-P. Goldman, A. Auchlin, S. Roekhaut, A. C. Simon, and M. Avanzi, "Prominence perception and accent detection in french. a corpus-based account." in *Speech Prosody*, 2010.

[19] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20(1), pp. 37–46, 1960.

[20] P. Mertens, "The prosogram: Semi-automatic transcription of prosody based on a tonal perception model," in *Speech Prosody*, 2004.

[21] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project," 2003.

[22] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Interspeech*, 2011.

[23] A. Lacheret-Dujour and F. Beaugendre, *La prosodie du français*. Paris: CNRS Editions, 1999.

[24] A. Séguinot, "L'accent d'insistance en français standard," *Studia Phonetica*, vol. 12, 1976.

[25] J.-P. Goldman, M. Avanzi, A. Lacheret-Dujour, A. C. Simon, and A. Auchlin, "A methodology for the automatic detection of perceived prominent syllables in spoken french," in *Interspeech*, 2007, pp. 98–101.

[26] J.-P. Goldman, M. Avanzi, A. Auchlin, and A. C. Simon, "A continuous prominence score based on acoustic features," in *Interspeech*, 2012.

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

[28] Hmm-based speech synthesis system (hts). [Online]. Available: http://hts.sp.nitech.ac.jp

[29] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation," in *International Conference on Spoken Language Processing (ICSLP)*, 1994, pp. 1043–1046.

[30] 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.