AUTOMATIC DETECTION AND CORRECTION OF SYNTAX-BASED PROSODY ANNOTATION ERRORS

Sandrine Brognaux[1], Thomas Drugman[2], Richard Beaufort[3]

[1] CENTAL and ICTEAM - Université catholique de Louvain, Belgium
[2] TCTS Lab - Université de Mons, Belgium
[3] Nuance Communications, Inc.*

ABSTRACT

Both unit-selection and HMM-based speech synthesis require large annotated speech corpora. To generate more natural speech, considering the prosodic nature of each phoneme of the corpus is crucial. Generally, phonemes are assigned labels which should reflect their suprasegmental characteristics. Labels often result from an automatic syntactic analysis, without checking the acoustic realization of the phoneme in the corpus. This leads to numerous errors because syntax and prosody do not always coincide. This paper proposes a method to reduce the amount of labeling errors, using acoustic information. It is applicable as a post-process to any syntax-driven prosody labeling. Acoustic features are considered, to check the syntax-based labels and suggest potential modifications. The proposed technique has the advantage of not requiring a manually prosody-labelled corpus. The evaluation on a corpus in French shows that more than 75% of the errors detected by the method are effective errors which must be corrected.

Index Terms: Prosody, Speech Synthesis, Annotation, Corpus

1. INTRODUCTION

Large speech corpora play a key role in both unit-selection and HMM-based speech synthesis. These corpora have to be annotated to provide information about the nature of each unit, usually phonemes or diphones. Information includes the position of these latter in the current sentence, word or syllable but also the part of speech of the carrier word, the structure of the carrier syllable, etc. The suprasegmental realization (duration, fundamental frequency and energy) of the unit has also to be considered. Several strategies have been proposed for this purpose.

The most straightforward way is to label each unit with its exact values of fundamental frequency (F0), duration and energy. This however implies that precise values should also be predicted for each phoneme of any new sentence to synthetize. This prediction is particularly difficult to make as it should correspond to a natural realization and be close to existing units in the database. Metrics should also be developed to compute a distance between prediction values and values available in the database, taking into account the relative relevance of each acoustic parameter, which is especially challenging.

To alleviate this problem, various annotation schemes have been proposed. In [1], it is proposed to automatically cluster similar units. These clusters can be accessed with a decision tree based on linguistic and acoustic criteria. Metrics to define the distance between target and database unit are easily computed as the distance between the unit and the centroid of its node. At synthesis time, however, some acoustic values still need to be predicted to browse the tree. In order to avoid that prediction, the use of symbolic information like 'tones', *i.e.* prosodic labels, has been presented by Campbell [2]. Various acoustic values are gathered into a same tone, that should reflect a general prosodic realization. This reduces the number of possible values to predict, while smoothing acoustic variations that might be related to a similar prosodic function. The tones are usually predicted on the basis of the syntactic structure of the sentence, as described below. In that respect, several symbolic labeling schemes can be exploited. Mertens' tones [3] indicate phrase-boundaries while ToBI [4] also assigns labels for prominence. Phrases can be defined as syntactic groups which *could* be isolated by means of pauses.

Conversely to [2], other speech synthesizers like the LiONs system presented in [5] do not rely on any prosodic label. This latter approach makes use of pure linguistic context only, like the position of the word in the sentence and its part of speech. It allows avoiding errors made by the tone prediction and assumes that syntactic parameters are enough to characterize the prosodic realization.

One of the goals of both [2] and [5] is to identify the location of phrase-end boundaries because they are often associated with major prosodic movements. To determine their position, punctuation is very helpful. However, phrases are not always delimited by punctuation marks. Most systems then rely on a heuristic segmentation called "chinks & chunks" [6] to identify phrases. It is based on a very simple rule which roughly assumes that there should be a boundary between a group of *content words* (chunks) and a group of *function words* (chinks). Here is an example of the application of this rule:

[There are several important changes][in the way][the quantifier rules][will work][for the remainder][of the course].

In this example, the last syllable of *changes*, *way*, *rules*, *work*, *remainder* and *course* will be assigned a specific prosodic label which should correspond to a specific prosodic realization.

The problem is that the application of this rule without any verification of the acoustic realization in the corpus is prone to errors. A first issue is that the simplistic principle of the "chinks & chunks" method does not allow it to take long distance dependencies, semantic information or specific prosodic rules into account. Among these rules is the "stress collision" [7], *i.e.* the fact that two consecutive stresses can only occur if they are divided by a boundary of higher rank. This boundary is usually marked by a melodic movement and syllabic lengthening and, possibly, the insertion of a pause between both stresses. The "chinks &

* The study was carried out while Richard Beaufort was still working at the CENTAL (Université catholique de Louvain, Belgium)

chunks", on the contrary, allows the presence of adjacent stresses, whatever their acoustic realization. A second issue is that, as shown in [8], prosodic boundaries do not always coincide with syntactic boundaries. Indeed, the speaker does not always produce a prosody which corresponds to the automatic syntactic analysis. Besides, it is well-known that the same sentence can be pronounced with various valid intonation patterns. These drawbacks of the algorithm lead to numerous labeling errors.

The application of the "chinks & chunks" on the following sentence gives this phrase segmentation:

[This is a ticket][to New York city].

However, in the corpus, the speaker might produce a slight pause after "to", to insist on the last group of the sentence. The segmentation would be the following:

[This is a ticket to][New York city].

The "chinks & chunks" algorithm, applied without taking the acoustic realization into account, proposes a labeling erroneous in two ways: [ticket] is erroneously labeled as phrase final, while [to] is not.

Incorrect prosodic labeling might induce a wrong choice of unit or acoustic model at synthesis time and affect the resulting speech quality. The inappropriate generated prosody can then be regarded by the listener as a "lack of understanding" of the sentence by the system. An optimal labeling of the units is thus crucial.

A way to solve this issue is to make use, at the labeling stage, of both acoustic and linguistic information to determine the phrase boundaries, as done in [9, 10, 11]. This, however, requires a prosody-annotated corpus to train the automatic annotator. Indeed, statistical models defining the acoustic values corresponding to each prosodic label should be learned on corpus. Ideally, the corpus should have been annotated manually. However, it is a time-consuming process which is impracticable for a large speech corpus. Besides, a different statistical model should be learned for each language. Conversely, the advantage of heuristic syntax-based algorithms like the "chinks & chunks" is that they do not require annotated corpora.

The goal of this paper is to propose a new solution which reduces the amount of errors made by syntax-based labeling. The innovation is that it is a post-processing method that can be adapted to any existing syntax-based labeling, while remaining automatic. It makes use of acoustic information to check the existing labels and to suggest modifications. The method focuses on phrase-end boundaries which are not associated with punctuation marks because they are prone to prosody labeling errors. The syllables and their corresponding phonemes are assigned a binary label which indicates if they are at the final boundary of a phrase or not. For the sake of clarity, these labels will be called 'tones' in the paper and be assigned two possible values: NB (not boundary) and B (boundary). However, the method could easily be adapted to apply to a more complex labeling scheme consisting of several different tones. The interest of the method is that it does not necessitate a manually prosody-annotated corpus since it is trained on the partly erroneous syntax-based labeling dataset. The proposed method has also been designed to be as generic as possible. Since the various parameters are trained on the target corpus, it is adaptable to any language.

This paper is organized as follows. The proposed method is first described in Section 2. It aims at identifying where the syntax-based labeling is potentially erroneous and deciding whether a correction is required or not. The results of the evaluation on a corpus in French are then shown in Section 3. In addition it is investigated how the use of acoustic information as a check on syntax-based labels can emphasize common errors made by syntax-driven prosody labeling algorithms. The errors found on a corpus in French, annotated with a classical "chinks & chunks", are further analyzed in Section 4. Finally Section 5 concludes and discusses further works.

2. DESCRIPTION OF THE METHOD

The proposed method relies on a speech corpus previously transcribed, phonetically-aligned and labelled in terms of parts of speech and prosody. The prosody labels must have been assigned at the syllable level and produced from syntax only. In this work, the initial prosodic labeling of the corpus is performed with the two following approaches (albeit the proposed method is transposable to any syntax-based prosody labeling technique):

- the classical "chinks & chunks" technique (CCBase) as described in [6],
- a modified version of the "chinks & chunks" proposed in [12] (CCModif). In this version, the grammatical categories are replaced by broader tags indicating if the word can or cannot be at the end or the beginning of a phrase. It then proposes a more complex rule which considers a longer sequence of words to determine the position of the phrase boundaries.

The general workflow of the proposed method is shown in Figure 1 and can be divided into three steps. In the first stage (Section 2.1), a set of acoustic features is extracted for each syllable of the speech corpus. An analysis of the feature distribution as well as their prediction ability highlights which of the considered features are the most relevant to discriminate between B and NB tones. In a second step (Section 2.2), these features and the initial tones (decided by the syntax-based labeling method) are used to train a decision tree. Then, each syllable is assigned a tone by the tree. This tone is compared to the initial syntax-driven one. Finally, if they are different, a third step (Section 2.3) determines, by means of an error score, whether this difference justifies a modification of the label or not.

To illustrate the different stages with effective results, the method is here applied on a corpus in French used for the LiONS unit-selection synthesizer [5]. This corpus is made of 3,339 isolated sentences read by a professional speaker for a total duration of 84 minutes. Each sentence corresponds to an independent audio file, its automatic syntactic analysis and automatically-extracted acoustic features (as achieved in Section 2.1). The whole corpus contains 22,261 syllables from which 10,774 (2,865 labelled as B and the rest as NB) were retained. It is worth noting that these syllables are neither followed nor preceded by a silence, initial and final syllables usually exhibiting specific acoustic features.

2.1. Extraction of Acoustic Features

The first step aims at extracting acoustic information from the sound files. For this, we here tried to consider a variety of features and to study their relevance to distinguish between both B and NB tones. Various features related to fundamental frequency (F0), duration and energy have been considered. These features can be mean values of

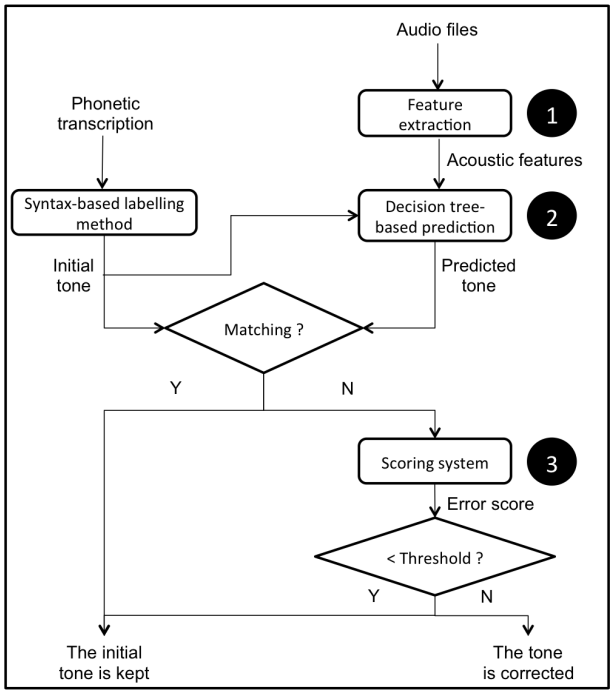

Fig. 1. *Workflow of the proposed method*

the syllable and its nucleus[1] but also delta values, *i.e.* comparisons with the value of the next/previous syllable and nucleus. Dynamic F0 values for the syllable and its nucleus can also be computed as well as prominence values. The "prominence value" is computed according to the formula in [14]. It was adapted to the max F0 and mean F0 ($\overline{F0}$) :

$$Prom(\overline{F0_i}) = \frac{\sum_{j=i-Ran(\overline{F0_i})}^{i+Ran(\overline{F0_i})} \overline{F0_i} - \overline{F0_j}}{2 * Ran(\overline{F0_i}) + 1} \quad (1)$$

where $Ran(\overline{F0_i})$ is the range of syllable i, *i.e.* the scope of the prominence in means of syllables on the left and on the right. The studied syllable is thus higher or lower than all the syllables within its scope. For example, "the F0 range is 0 if the syllable is neither a minimum, nor a maximum, and 2 if its value dominates those of the 2 syllables on its right and on its left" [14].

All values are normalized in z-scores, which allows comparing them (which will be notably required in Section 2.3). The z-score of a value measures its distance with respect to the mean value, in proportion to the standard deviation of this value. The z-score of the mean F0 of the i^{th} nucleus ($\overline{F0_i}$), for example, is computed as follows:

$$Zscore(\overline{F0_i}) = \frac{\overline{F0_i} - \overline{F0}}{s_{F0}} \quad (2)$$

where $\overline{F0}$ is the mean value of all the ($\overline{F0_i}$) and s_{F0} is its standard deviation. For the duration, each nucleus must be normalized with respect to the duration of the corresponding phoneme. This choice relies on the fact that the nature of the phoneme clearly affects its duration [15, 16].

[1]the nucleus is considered to be the intensity and sonorance peak of the syllable, *i.e.* usually its vocalic part [13].

The complete list of extracted features to test our method is shown in Table 1. No energy values were considered because the energy of the sound had been normalized during the recording stage.

The relevance of the various features was observed for the corpus in French labeled with the CCBase. As previously mentioned, its syntax-based labeling is partly erroneous. However, the amount of errors is slight enough to allow outlining general tendencies. In order to determine which features are the most relevant for the distinction between both B and NB tones, we first analyzed a measure of the distance between the feature distributions for these two classes. For this, the Kullback-Leibler (KL) divergence, known to measure the separability between two discrete density functions A and B, can be used [17]:

$$D_{KL}(A, B) = \sum_i A(i) \log_2 \frac{A(i)}{B(i)} \quad (3)$$

Since this measure is non-symmetric (and consequently is not a true distance), its symmetrised version, called Jensen-Shannon (JS) divergence, is often preferred. It is defined as a sum of two KL measures [17]:

$$D_{JS}(A, B) = \frac{1}{2} D_{KL}(A, M) + \frac{1}{2} D_{KL}(B, M) \quad (4)$$

where M is the average of the two distributions ($M = 0.5 * (A + B)$). In other words, the higher the JS divergence, the better the distributions are separable and the most relevant the feature is for discriminating between B and NB tones. Table 1 indicates the values of the JS divergence for the features considered in this work. These results allow to draw a first conclusion regarding the relative relevance of the features.

It should be highlighted that some features seem to draw a significantly better distinction between B and NB tones than others. F0 seems to play a bigger role in the distinction than the duration (with a JS average value of 0.0526 compared to 0.0205 for duration features). Prominence values related to F0 lead to the highest JS divergences. When analyzing the distribution of the mean F0 prominence value on the corpus, we notice that 49% of the NB syllables are prominent compared to 67% of the B syllables. Most duration values are assigned a much lower JS divergence which indicates their weaker discriminatory power.

The most discriminant acoustic features can also be determined by a decision tree with a greedy algorithm launched on a balanced corpus. Classically, the building of the tree considers all features to choose the criterion to split a node. The stepwise function modifies this behavior: it limits the number of characteristics that are taken into account. At each step, the system analyses which feature allows, when added to the other features already chosen at previous steps, to maximize the prediction rate. It gradually increments the number of exploited features. This offers a ranking of the features according to their discriminatory power in terms of tones. Conversely to the JS divergence values which indicate the intrinsic performance of each feature individually, the stepwise takes the complementarity of the various features into account. The application of this algorithm on the corpus also proved the predominance of F0, with the three best descriptors being F0-related features (see Table 2). Here again, the first place is occupied by the prominence mean F0 which shows the better discriminatory power of that feature. In the fourth place, a duration characteristic complements the information given by the three first features. The prediction gain, however, is rather low (1.15%).

The duration is then observed to play a minor role in the distinction between both B and NB tones.

Table 1. List of extracted acoustic features together with their Jensen-Shannon divergences

Features	JS divergence
Duration features	
Nucleus duration	0.0248
Syllable duration	0.0117
Delta of duration with previous nucleus	0.0130
Delta of duration with next nucleus	0.0333
Delta of duration with previous syllable	0.0102
Delta of duration with next syllable	0.0428
Prominence of nucleus duration	0.0130
Prominence of syllable duration	0.0148
F0 features	
Max F0 of nucleus	0.0457
Mean F0 of nucleus	0.0462
Delta of max F0 with previous nucleus	0.0529
Delta of max F0 with next nucleus	0.0471
Delta of mean F0 with previous nucleus	0.0605
Delta of mean F0 with next nucleus	0.0514
Delta of F0 between beginning and end of nucleus	0.0289
Prominence of max F0 of nucleus	0.0643
Prominence of mean F0 of nucleus	0.0761

Table 2. Acoustic features selected by the greedy algorithm (stepwise) and the respective correct prediction rate of the tree

Step	Feature(s) selected	Correct prediction rate by the decision tree
1	Prominence of mean F0 of nucleus	64.52%
2	+ Delta of mean F0 with next nucleus	66.72%
3	+ Mean F0 of nucleus	68.94%
4	+ Delta of duration with next syllable	70.09%
5	+ Delta of mean F0 with prev. nucleus	70.38%

2.2. Prediction of a Tone with a Decision Tree

The second step is the prediction of a tone by a decision tree. To assign a label to each syllable, a decision tree (Wagon CART [18]) is trained on the corpus with the acoustic features extracted at step one (see Table 1) and the initial syntax-driven tones. As previously mentioned, the syntax-driven labeling is partly erroneous. However, the amount of errors made by the syntax-driven prosody annotation algorithm is slight enough to allow for the training of the tree. A rather high stop value is fixed (here, 50) to prevent the tree from modeling the errors. This means that a node is not split if it contains less than 50 occurrences in the training dataset. In turns, 90 % of the corpus is used to train the tree that assigns a tone to each syllable of the reminding 10% of the corpus. This is done iteratively to predict tones for all the syllables of the corpus. Finally, the predicted label is compared to the initial syntax-based label. If they are identical, the initial label is assumed to be correct and is kept. If not, a further analysis is performed in Section 2.3 to determine whether the label should be changed. In other words, it is assumed at this stage that, if the proposed method and the syntax-based labeling technique give the same tone decision, this latter decision is highly likely to be correct (which was corroborated through our informal observations).

2.3. Scoring System

The third and final step is the scoring system. This stage assigns an Error Score (ES) to each syllable for which the syntax-based label differs from the label predicted by the tree. The aim is to determine whether the tone should actually be changed or if the syllable is a borderline case. A high ES means that the syllable is a real outlier with the syntax-based label and that it is very likely that it should be labeled with the other tone.

This step can be subdivided into two stages:

- A new regression tree is first used to predict the value of a pre-selected set of acoustic features. This new tree is trained on the initial syntax-driven labeling. Conversely to the first decision tree, it is based on the linguistic features and the syntax-based tone only, and not on acoustic features which are predicted. Various linguistic features can be considered, as shown in Table 3.
- The distance between the prediction of these features and their realization is calculated to obtain a score based on the following formula:

$$ES = \frac{1}{N} \sum_{i=1}^{N} D(R(F_i), P(F_i)) \quad (5)$$

where $D(R(F_i), P(F_i))$ is the Euclidian distance expressed in z-score which measures how much the acoustic realization $(R(F_i))$ of feature i (from the pre-selected set of N features) is close to its prediction $(P(F_i))$. Because of the use of z-score measures for all features, scale factors are avoided and all scores can be compared, whatever acoustic features are pre-selected. In other words, the measure indicates the average z-score difference between several predicted acoustic features of the syllable and their realization.

In the following, we only considered the two best features from Table 2 as the complementary information brought by the other features was found to be only minor.

The scoring system was applied to the problematic syllables of the corpus in French. The linguistic features considered to train the regression tree are shown in Table 3. The potential labeling errors pointed at by our method were then analyzed. Informal perceptual analyses, based on the listening of several hundreds of syllables for which the predicted label differed from the syntax-based initial one, were performed. It indicated that syllables with ES higher than 1 are often errors, meaning that the label should be changed.

The percentage of syllables for which our method assigned a tone different from the initial syntax-based tone is shown in Table 4. An insightful observation regards the lower percentage of divergences between the tone predicted by our method and the tone assigned by the modified version of the "chinks and chunks" (CC-Modif). This seems to indicate that this version solves some problems experienced by the classical "chinks and chunks". It should be noted that less than 2% of the syllables are assigned a high ES which indicates that their label is erroneous and should be corrected. This percentage of potential errors present in the corpus is fairly low but sufficient to produce unwilling prosodic movements in the synthetic speech. The correction of these errors should improve the quality of the generated voice.

Table 3. Linguistic features considered for the training of the regression tree to compute the ES

Classes	Values
Part of speech of current word	Infinitive/noun/adjective/etc.
Part of speech of previous word	Infinitive/noun/adjective/etc.
Part of speech of next word	Infinitive/noun/adjective/etc.
Phoneme of the nucleus of the syllable	o/a /i/etc.
Type of phoneme of the nucleus of the syllable	NasalVowel/OralVowel
Structure of the syllable	CVC/CV/etc.
Type of phoneme after the nucleus	Liquid/Fricative/Plosive/etc.

Table 4. Percentage of potential errors pointed at by our method for the syntax-based labels (with CCBase and CCModif) of the corpus in French

Chinks&Chunks labels	All potential errors	Potential errors with ES > 1
CCBase	24.59%	1.52%
CCModif	20.70%	0.97%

3. EVALUATION

3.1. Evaluation Protocol

Five human expert annotators were asked to label 50 syllables for which the CCBase syntax-based label differed from the label predicted by the proposed method. They assigned each syllable a "prominent/boundary" label or not. We decided to include both terms in the label. Indeed, the acoustic base of our system detects more, in this view, prominent syllables, even if it is used to correct boundary labels. This group of syllables consisted of 25 divergences of each type, *i.e.* syntax-based tone is B and tone predicted by our method is NB; and the opposite. The syllables were chosen as those with the highest ES. They were all syllables with ES higher than 1.15 (between 1.17 and 3.94). Each annotator could also label a syllable "?" if its prosodic nature seemed doubtful.

3.2. Results

A high inter-annotator agreement (IAA) was obtained : 87.98% when not taking the "?" into account. This rate was obtained as follows:

$$IAA = \frac{1}{n}\sum_{i=0}^{n}\frac{max(NB_i, B_i)}{n_a} \quad (6)$$

where n_a is the number of human annotators (*i.e.* 5), n is the total number of syllables to label (*i.e.* 50) and NB_i and B_i are the number of human annotations in which syllable i is labelled NB and B, respectively.

The agreement between the human-made labeling and the correction proposed by our method is shown in Table 5. An average agreement rate of 78.72% is noticed between the annotators and the corrections proposed by our method. This means that more than 3/4 of the syllables pointed at by the method as erroneously labelled should indeed be changed.

This rate could even increase because of its distribution between the two error types. It should be reminded that the French corpus is unbalanced in terms of tones. The a priori probability of NB tones is significantly higher (0.734). This explains why Table 5 indicates that NB syllables considered as B by the method seem to be more often effective errors than the opposite. Syllables are only labelled B by the tree in step 2 if they truly display boundary characteristics. The corpus was not balanced at this stage because it would strongly lower the number of occurrences to train the tree. It implies that the agreement between the annotators and the system could still increase with larger balanced corpora at disposal.

Table 5. Percentage of right corrections proposed by our system (for CCBase tones) for 50 syllables with ES higher than 1.15

Proposed correction	Percentage of effective errors detected by our method
NB –> B	91.67%
B –> NB	65.22%

4. QUALITATIVE ANALYSIS OF COMMON SYNTAX-BASED LABELING ERRORS

A qualitative analysis of the errors detected by the method on the corpus in French brings out common errors made by the classical "chinks & chunks" (CCBase). We try here to list some generic cases that might also be applicable to other languages. The illustrations are given in French and should be of interest for French language experts.

Among syllables erroneously labelled NB can be found:

1. *Cases of emphasis*: the "chinks & chunks" algorithm does not consider emphatic stresses. Even if less present in neutral speech, they can still be found on specific words. In the corpus, they were found on contrast words (e.g."toutefois") and stressed pronouns (e.g."eux").
2. *Numbers*: in the simple "chinks & chunks" used here, no boundary is identified between two numbers. However, there is often a boundary in longer numbers such as "mil-neuf-cent-quatre-vingt".
3. *Boundary between verbs and negations*: the French negative adverbs "ne" and "pas" are processed in the same way by most "chinks & chunks" algorithms. However, there is generally no boundary between the verb and "pas" (linked to the verb) but well between the verb and "ne" (linked to the following verb). Ex: "Les entreprises dans lesquelles la société investit ne sont pas jugées sur leurs seules performances économiques." This could be corrected in the "chinks & chunks" algorithm by changing the categories corresponding to the respective words.

For syllables erroneously labelled as boundaries, the following cases can be brought out:

1. *Elements that make up a whole*: some groups of words are not considered as making up a whole and are assigned a boundary label in the middle (*e.g.* années 80).
2. *Stress "collision"*: as previously mentioned, [7] showed that two consecutive stresses can only occur if divided by an intonative boundary of higher rank. There is thus no boundary in : "Quelques blagues fusent, mal assurées." because of the stress of higher rank on "fusent".

3. *Items after "avoir" and "être"*[2]: although there is usually a boundary between the verb and the complement, it does not seem to be often the case with these two auxiliaries (*e.g.* "J'aurais dû avoir sept ou huit enfants.")

In general, many erroneously-labelled syllables are found in ill-structured or ambiguous sentences that seem to confuse the speaker. Grammatically-correct sentences with complementary information (pronunciation, liaisons, pauses, etc.) should be given to the speaker.

5. CONCLUSION AND PERSPECTIVES

Most text-to-speech systems available on the market rely on the use of a large speech corpus. To produce a more natural voice, a good prosody labeling of the phonemes or diphones of the corpus is needed. Usually, the labels are determined on the sole basis of syntax, without checking the acoustic realization in the sound. The discrepancy existing between prosodic and syntactic boundaries often leads to numerous errors.

This paper proposed an automatic method to detect and correct these errors. It can be applied as a post-processing method to any syntax-driven prosody annotation. It makes use of acoustic information to check the syntax-based labels and determines those which should be changed. In other words, it helps identifying acoustic outliers among a set of syllables labeled with the same prosody label. The interest of the proposed method is to be adaptable to any existing syntax-based prosody labeling and to any language. Besides, it does not require any manually prosody-annotated corpus to train the system. The originality of the method is that it uses the partly erroneous syntax-based prosody labeling to train a system that will isolate potential errors. The proposed approach allows a quick correcting of existing corpora. Experiments on a corpus in French showed that more than 75% of the potential errors detected by the method are effective errors that ought to be corrected. The correction of these errors is bound to improve the prosody of the generated speech. Further research will focus on the resulting improvement in terms of synthesized speech quality. It would be insightful to assess the impact produced by the corrections made in the prosody labeling.

The method also brings out typical flaws of syntax-based prosody labeling techniques. This paper showed, for example, that the widely-used "chinks & chunks" algorithm tends to treat erroneously compound phrases and to annotate adjacent stresses that are not realized. Future research should study if specific modifications of this algorithm would help to improve the prosody labeling.

6. ACKNOWLEDGEMENTS

Sandrine Brognaux is supported by the "Fonds National de la Recherche Scientifique" (FNRS). Authors would like to thank M. Avanzy, P. Mertens, S. Roekhaut and A.-C. Simon who put their expertise at the service of this study by labeling the 50 syllables used in the evaluation. They are also grateful to Vincent Pagel for his insightful ideas regarding this study.

7. REFERENCES

[1] A. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *Proc. of Eurospeech 97*, 1997, pp. 601–604.

[2] W.N. Campbell, "Processing a speech corpus for CHATR synthesis," in *Proc. of the International Conference on Speech Processing*, 1997, pp. 183–186.

[3] P. Mertens, *L'intonation du français. De la description linguistique la reconnaissance automatique.*, Ph.D. thesis, Univ. Leuven (Belgium), 1987.

[4] 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 *Proc. of ICSLP*, 1992, pp. 867–870.

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

[6] M.Y. Liberman and K.W. Church, "Text analysis and word pronunciation in text-to-speech synthesis," in *Advances in Speech Signal Processing*, Sadaoki Furui and M. Mohan Sondi, Eds., pp. 791–831. Dekker, New York, 1992.

[7] F. Dell, "L'accentuation dans les phrases en français," in *Forme sonore du langage*, F. Dell, D. Hirst, and J.R. Vergnaud, Eds., pp. 65–122. Hermann, Paris, 1984.

[8] L. Degand and A. C. Simon, *Where Prosody Meets Pragmatics: Research at the Interface*, chapter Mapping Prosody and Syntax as Discourse Strategies: How Basic Discourse Units Vary Across Genres, pp. 79–105, Studies in Pragmatics. Emerald, 2009.

[9] N. Braunschweiler, "The prosodizer : Automatic annotations of speech synthesis databases," in *Proc. of Speech Prosody 2006*, Dresden (Germany), 2006.

[10] A. Wagner, "Analysis and recognition of accentual patterns," in *Proc. of Interspeech 2009*, 2009, pp. 2427–2430.

[11] V. R. Sridhar, S. Bangalore, and S. Narayanan, "Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 4, pp. 797–811, May 2008.

[12] R. Beaufort, *Application des Machines à Etats Finis en Synthèse de la Parole. Sélection d'unités non-uniformes et Correction orthographique*, Ph.D. thesis, FUNDP, Namur, 2008.

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

[14] V. Pagel, *De l'utilisation d'informations acoustiques suprasegmentales en reconnaissance de la parole continue*, Ph.D. thesis, Université Henry Poincaré, Nancy, 1999.

[15] H. A. Rositzke, "Vowel-length in general american speech," *Language*, vol. 15, pp. 99–109, 1939.

[16] A. Di Cristo, *De la microprosodie à l'intonosyntaxe*, Ph.D. thesis, Université de Provence, Aix-en-Provence, 1985.

[17] J. Lin, "Divergences measures based on the shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, pp. 145–151, 1991.

[18] P. Taylor, R. Caley, A. W. Black, and S. King, "Wagon CART building program," in *Edinburgh Speech Tools Library, System Documentation Edition 1.2*. 1999, Centre for Speech Technology, University of Edinburgh, `http://festvox.org/docs/speech_tools-1.2.0/x3475.htm` [accessed 20-August-2009].

[2] "to have" and "to be", respectively