# SLAM:
# Automatic Stylization and Labelling of Speech Melody

*Nicolas Obin*[1]*, Julie Beliao*[2]*, Christophe Veaux*[3]*, Anne Lacheret*[2]

[1] IRCAM - UMR STMS IRCAM-CNRS-UPMC, Paris, France
[2] MoDyCo - UMR 7114, Université de Paris Ouest, Nanterre, France
[3] Centre for Speech Technology Research, Edinburgh, UK

## Abstract

This paper presents SLAM : a simple method for the automatic Stylization and LAbelling of speech Melody. This main contributions over existing methods are : the alphabet of melodic contours is fully data-driven, an explicit time-frequency representation is used to derive complex melodic contours, and melodic contours can be determined over arbitrary prosodic/syntactic units. Additionally, the system can handle some specificities of spontaneous speech (e.g., multi speakers, speech turns and speech overlaps). A preliminary experiment conducted on 3 hours of spoken French indicates that a small number of contours is sufficient to explain most of the observed contours. The method can be easily adapted to other stressed languages. The implementation is open-source and freely available [†].

**Index Terms** : intonation, stylization, automatic labelling, prosody, syntax.

## 1. Introduction

The transcription of speech prosody aims at representing the variations of speech prosody that are considered as relevant with an alphabet of elementary symbols [1, 2, 3, 4], which each instantiates a function in the speech communication process. The inventory of this alphabet is desired to facilitate further studies on the function of speech prosody in speech communication, from prosody/syntax to prosody/discourse and dialogue interfaces, from formal to spontaneous speech.

The first representation of French intonation in terms of global contours [5] focused on modal intonation : a global melodic contour specifies the modality of a sentence (e.g., interrogation, exclamation). More recently, this paradigm was extended to the representation of intonation as the superposition of melodic contours over various syntactic units [6]. The main contribution of this paper tends to the generalization of this paradigm to any linguistic unit – prosodic and syntactic. In other words, we assume that a specific dictionary of elementary contours can be derived for each linguistic unit.

This paper presents a novel method for the automatic labelling of melodic contours over arbitrary prosodic/syntactic units. The main contribution of the method compared to existing methods [2, 3, 7, 8, 9] can be summarized as follows :

- The melodic system (i.e., the alphabet of melodic contours) is fully data-driven (bottom-up processing).
- An explicit time-frequency representation is used to describe complex melodic contours.
- The proposed representation handles a large variety of prosodic/syntactic units : from local (e.g., syllable) to global contours (here, prosodic and syntactic).

Additionally, the representation is normalized with respect to the average range of a speaker, and handles some particularities of spontaneous speech (e.g., multi speakers, speech turns, speech overlaps). Lastly, the implementation is open-source and freely available.

The remainder of this paper presents the main principles of the method used for the transcription of melodic contours over arbitrary prosodic/syntactic units. The compact form of the alphabet of contours proves the efficiency of the proposed method : around 10/20 elementary contours suffice to explain 95% of the observed contours.

## 2. Intonation Labelling

### 2.1. Speech Preprocessing

The only external requirements for the automatic labelling are : the estimation of the fundamental frequency of speech (F0), and the segmentation of speech into speech units that are desired for the description of speech prosody (arbitrary prosodic/syntactic units). For the F0 estimation, popular methods are freely available (e.g., STRAIGHT [10], YIN [11], and SWIPE [12]). Also, many refinements to the F0 estimation exist to facilitate further processing - from F0 periodicity estimation (and voiced/unvoiced decision - similarly to [8]), to F0 smoothing and interpolation methods. For the speech segmentation, speech-to-text alignment methods exist (IRCAMALIGN [13], EASYALIGN [14], and SPPAS [15] among others - usually based on the open-source HTK library [16]) for the segmentation of speech into phonemes, syllables, words, and phrases. Also, alternative methods exist for language-independent speech segmentation into syllables and phrases [17].

### 2.2. Acoustic Representation

The acoustic stylization of speech melody consists in representing the F0 variations that are considered as relevant for the description of speech prosody. In general, F0 stylization methods are based on the representation of F0 variations according

†. This study was supported by the French National Research Agency (ANR) for the RHAPSODIE project : reference prosody corpus of spoken French. The resource is implemented in python and freely available on : `https://github.com/jbeliao/SLAM/`. The current release supports PRAAT TextGrid input/output format for segmentation and labelling. There is no need for preliminary F0 estimation, which is processed automatically with the python implementation of the SWIPE algorithm ( see `https://github.com/kylebgorman/swipe/` for details).

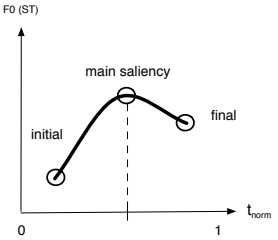


FIGURE 1 – *Acoustic representation of a contour.*

to a small set of parameters that are used to describe slowly time-varying F0 variations [18, 19, 20, 21]. In the proposed method, the F0 contour is represented by a set of 5 acoustic values for each given unit :

1. INITIAL : the initial value of the F0 on the unit. This value corresponds to the first F0 value for which the acoustic frame is considered as voiced ;
2. FINAL : the final value of the F0 on the unit. This value corresponds to the last F0 value for which the acoustic frame is considered as voiced ;
3. MAIN SALIENCY : the value corresponding to the most salient F0 peak – if one exists. The F0 variations over a unit can be decomposed into a main saliency (optional, if present) and a set of secondary salience (optional, if present). The method assumes that only the main saliency contributes to the definition of a global contour, while secondary salience contribute to the internal structure of the global contour, and can be neglected in a first-order approximation.

Finally, the following values are added to the description :

4. MAIN SALIENCY POSITION : the time position of the main saliency ;
5. LOCAL REGISTER : the mean F0 over the unit.

All frequency values are expressed in semi-tones (STs), with respect to the overall mean F0 of the speaker :

$$\text{F0[ST]} = 12 \times \log_2 \frac{\text{F0[Hz]}}{\text{F0}_{\text{mean}}\text{[Hz]}} \quad (1)$$

All time positions are expressed relative to the boundaries of the unit :

$$t_{norm} = \frac{t - t_{start}}{t_{end} - t_{start}} \quad (2)$$

This acoustic representation adapts automatically to the nature of the prosodic/syntactic unit. Here, the notion of micro and macro prosodic variations is assumed to be relative to the linguistic unit considered : for short units (and local contours ; e.g., syllable), the phoneme variations will be considered as micro variations compared to the syllable variations ; for large units (and global contours ; e.g., phrases), the syllable variations will be in turn considered as micro variations compared to the larger unit.

### 2.3. Symbolic Representation

The acoustic representation presented in the previous section serves as a time-frequency representation for the labelling of contours.

#### *2.3.1. Frequency Quantization*

First, frequency values are represented with respect to 5 pitch levels covering the whole F0 range of the speaker (table 1). Each pitch level covers a range of 4 semi-tones centred on the average F0 value of the speaker. For instance, the medium range covers from - 2 STs to + 2 STs around the average range of the speaker ; the high range covers from +2 STs to +6 STs ; and the extreme-high range covers all values that exceed +6 STs.

| PITCH LEVELS | DESCRIPTION | RANGE (STs) |
|---|---|---|
| H | extreme-high | $>$ +6 |
| h | high | +2/+6 |
| m | medium | -2/+2 |
| l | low | -2/-6 |
| L | extreme-low | $<$ -6 |

TABLE 1 – *Pitch levels used for the symbolic representation.*

From this representation the sequence of initial/final/saliency values can be converted into a corresponding sequence of pitch levels. Then, this representation can be used to describe static tones, simple contours, and complex contours. An illustration is provided in figure 2.

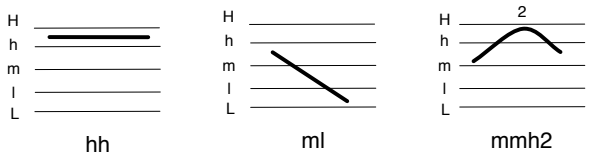


FIGURE 2 – *Types of contours that can be described. From left to right : a static tone (flat contour in the high range of the speaker), a simple contour (a falling contour from the medium to the low range of the speaker), and a complex contour (medium to medium with a saliency observed in the high range of the speaker in the middle part of the unit).*

Additionally, the main saliency is considered as significant only if the corresponding point differs by more than 2 ST from the initial and the final points. If this is not the case, the main saliency is not considered as relevant, and is removed from the symbolic representation of the contour. An illustration is provided in figure 5.

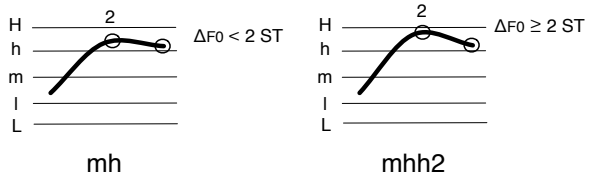


FIGURE 3 – *Determination of the presence of a saliency. On the left, the saliency is not considered as relevant, and the contour is transcribed as "mh" (medium to high) ; on the right, the saliency is relevant, and the contour is transcribed as "mhh2" (medium to high with the presence of a saliency in the second part of the unit).*

*2.3.2. Time Quantization*

Second, the time position of the main saliency is represented with respect to 3 time positions, which are determined from the relative position of the saliency within the unit and the decomposition of the unit into 3 equal parts 2. An illustration of contours with various positions of the main saliency is provided in figure 5.

| TIME POSITION | MAIN SALIENCY |
|---|---|
| 1/3 | first part of the unit |
| 2/3 | middle part of the unit |
| 3/3 | last part of the unit |

TABLE 2 – *Time position of the main saliency of a contour used for the symbolic representation.*

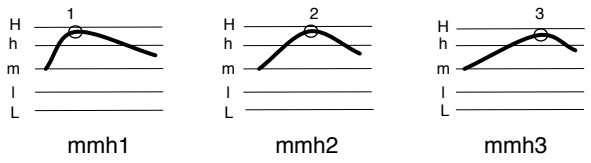


FIGURE 4 – *Contour labelling with regard to the position of the saliency within the unit. From left to right, a saliency is observed in the first/middle/last part of the unit.*

*2.3.3. Formal Representation*

Finally, the time-frequency representation is formally described as :

Initial Final [Saliency] [Position]

where [.] indicates optional fields dependent on the relevancy of the main saliency. Illustrations of the various contours that can be represented are shown in figure 2.

The alphabet of this formal representation is expressive in the sense that the alphabet can cover a large variety of contours - here, 400. Theoretically, the optimality of a phonological system assumes the smallest number of elements required to account for the largest number of observations. Practically, a compact alphabet is desired to facilitate labelling and linguistic interpretation. For instance, the inventory in the ToBI system (in order to describe phrasal tones and pitch accents) includes two elementary tones (H, L), and around 5 to 8 pitch accents have been proposed for American-English [22]. In order to address the efficiency of the proposed representation, a preliminary experiment will be described in section 3, which proves that a small number of contours (around 10/20) is sufficient to explain most of speech prosody variations in the real-conditions of ordinary speech. An illustration of the acoustic representation and the labelling of contours is provided in figure 5.

## 3. Experiment

A preliminary experiment was conducted on the RHAPSODIE treebank (prosody/syntax) of spoken French [23], composed of speech recordings of French ordinary speech and orthographic transcription. The transcription and the annotations are all aligned on the speech signal : phonemes, syllables, words, speakers, speech turns, speech overlaps. The RHAPSODIE treebank comprises : 57 speech recordings, 3 hours of ordinary speech, and 33,000 words ; multiple situations : monologue/dialogue, formal/informal ; multiple speakers : male/female.

Firstly, the analysis of contours reveals that a small number of contours suffices to explain most of the observed contours : around 10/20 elementary contours suffice to explain 95% of the observed contours, regardless of the prosodic/syntactic unit (table 3, see [23] for details). This can be interpreted as follows : a small number of elementary contours commonly serves for usual speech communication, and a variety of rare contours may convey specific speaker and/or expressive information (e.g. emotions). This constitutes a first validation concerning the acoustic/symbolic representation : the representation is efficient (a small alphabet explains most of the observations) and expressive (the representation can describe a variety of contours that are not accounted by the standard alphabet).

Secondly, the distribution of the most observed contours is detailed for some prosodic/syntactic units in figure 6 (here, syllable, discourse markers, and illocutionary units). In particular, the alphabet substantially changes depending on the prosodic/syntactic unit : in comparison with the common syllable unit, the discourse marker and illocutionary units present a larger variety of contours (e.g., extreme ranges, complex contours), that potentially instantiates various functions : from modalities, to semantic and pragmatic. This constitutes a second validation concerning the labelling of contours over various prosodic/syntactic units : an alphabet of contours can be derived specifically for each prosodic/syntactic unit.

In conclusion, the representation can derive a small alphabet of contours that is specific to each prosodic/syntactic unit, that can be advantageously used for automatic labelling. This is crucial for further research on the role of prosody in speech communication : the grail search for the mapping of forms and functions.

## 4. Conclusion

This paper presented a simple method for the automatic labelling of intonation. The proposed method presents various advantages over existing methods : the alphabet of contours is fully data-driven, an explicit time-frequency representation is used to derive complex contours, and contours can be determined over arbitrary prosodic/syntactic units. A preliminary experiment conducted on 3 hours of spoken French indicates that a small number of contours is sufficient to explain most of the observed contours. The method can be easily adapted to other stressed languages. The implementation is open-source and freely available. This representation will be further used to study the role of speech prosody in speech communication, from prosody/syntax and prosody/discourse interfaces [24, 25], to the modelling of speech prosody for text-to-speech synthesis and voice conversion [26].

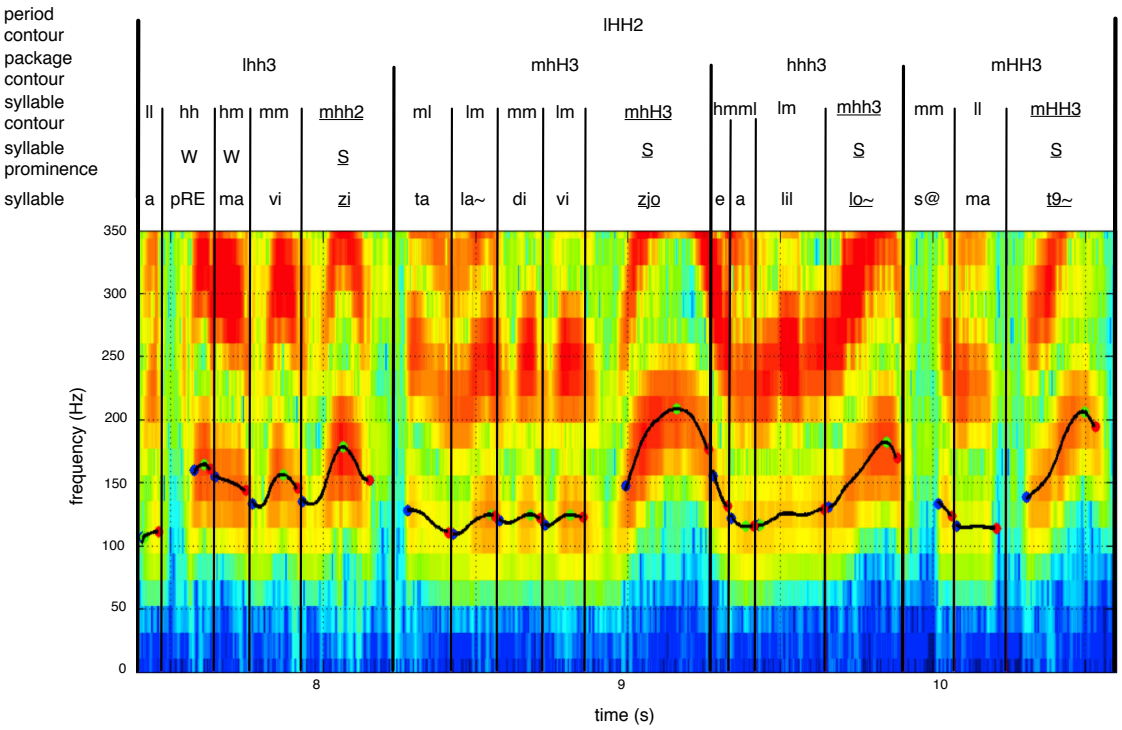


FIGURE 5 – *Acoustic representation and labelling of contours over syllable, prosodic package, and prosodic period for the speech sequence : "Après ma visite à Landivisiau et à l'île Longue ce matin" ("After my visit to Landivisiau and l'île Longue this morning"). Speech sample [Rhap-M2001, corpus C-PROM] : monologue, ordinary speech. Blue and red dots denote initial and final values, respectively ; and green dots intermediate saliencies. For syllable prominence : W indicates a weak prominence, and S a strong prominence. Information about the last syllables of a prosodic package are underlined.*

| PROSODIC UNITS | | # CONTOURS (> 95%) | SYNTACTIC UNITS | | # CONTOURS (> 95%) |
|---|---|---|---|---|---|
| syllable | (43192) | 8 | discourse marker | (966) | 7 |
| word | (32083) | 9 | pre-kernel | (855) | 15 |
| foot | (22705) | 11 | post-kernel | (158) | 39 |
| group | (18104) | 12 | integrated-kernel | (142) | 23 |
| package | (14206) | 15 | illocutionary unit | (2847) | 28 |
| period | (2507) | 29 | ... | ... | ... |

TABLE 3 – *Occurrence of the contours observed over the prosodic/syntactic units. From left to right : nature of the prosodic/syntactic unit, total number of units observed, and number of contours that explain 95% of of the observed contours for each unit.*

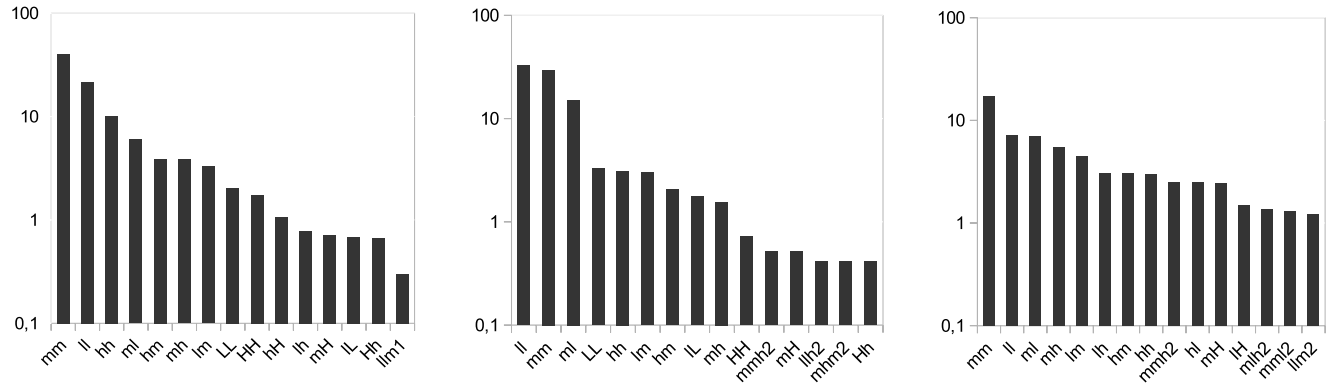


FIGURE 6 – *Proportion of the 15 most frequent contours observed for a set of prosodic/syntactic units (log % of occurrences). From left to right : syllable, discourse marker, and illocutionary units.*

## 5. References


[1] 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 of Spoken Language Processing*, Banff, Canada, 1992, pp. 867–870.

[2] P. Taylor, "The Rise/Fall/Connection Model of Intonation," *Speech Communication*, vol. 15, pp. 169–186, 1994.

[3] E. Campione, D. Hirst, and J. Véronis, *Automatic Stylisation and Symbolic Coding of F0 : Implementations of the INTSINT Model*. Dordrecht : Kluwer, 2000, ch. Intonation. Research and Applications.

[4] B. Post, E. Delais-Roussarie, and A.-C. Simon, "IVTS, un Système de Transcription pour la Variation Prosodique," *Bulletin de la Phonologie du Français Contemporain*, vol. 6, pp. 51–68, 2006.

[5] P. Delattre, "Les Dix Intonations de Base du Français," *The French Review*, vol. 40, no. 1, pp. 1–14, 1966.

[6] V. Aubergé, "La Synthèse de la Parole : "Des Règles aux Lexiques"," PhD. Thesis, Université Pierre Mendès-France, Grenoble, France, 1991.

[7] K. Syrdal, A., J. Hirschberg, J. McGory, and M. Beckman, "Automatic ToBI Prediction and Alignment to Speed Manual Labeling of Prosody," *Speech Communication*, vol. 33, no. 1-2, pp. 135–151, 2001.

[8] P. Mertens, "The Prosogram : Semi-Automatic Transcription of Prosody based on a Tonal Perception Model," in *Speech Prosody*, Nara, Japan, 2004, pp. 549–552. [Online]. Available : http://bach.arts.kuleuven.be/pmertens/prosogram/

[9] ——, "Automatic Labelling of Pitch Levels and Pitch Movements in Speech Corpora," in *Tools and Resources for the Analysis of Speech Prosody (TRASP)*, Aix-en-Provence, 2013.

[10] H. Kawahara, H. Katayose, A. De Cheveigné, and R. D. Patterson, "Fixed Point Analysis of Frequency to Instantaneous Frequency Mapping for Accurate Estimation of F0 and Periodicity," in *Eurospeech*, Budapest, Hungary, 1999, pp. 2781–2784.

[11] A. De Cheveigné and H. Kawahara, "YIN, a Fundamental Frequency Estimator for Speech and Music," *Journal of the Acoustic Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[12] A. Camacho, "SWIPE : A Sawtooth Waveform Inspired Pitch Estimator for Speech and Music," PhD. Thesis, University of Florida, 2007. [Online]. Available : http://www.cise.ufl.edu/~acamacho

[13] P. Lanchantin, A. Morris, X. Rodet, and C. Veaux, "Automatic Phoneme Segmentation with Relaxed Textual Constraints," in *International Conference on Language Resources and Evaluation*, Marrakech, Morroco, 2008, pp. 2403–2407.

[14] J.-P. Goldman, "EasyAlign : a Semi-Automatic Phonetic Alignment Tool under Praat," in *Interspeech*, Florence, Italy, 2011. [Online]. Available : http://latlntic.unige.ch/phonetique

[15] B. Bigi and D. Hirst, "SPeech Phonetization Alignment and Syllabification (SPPAS) : a Tool for the Automatic analysis of Speech Prosody," in *Speech Prosody*, Shanghai, China, 2012. [Online]. Available : http://aune.lpl.univ-aix.fr/~bigi/sppas

[16] S. Young, "The HTK Hidden Markov Model Toolkit : Design and Philosophy," *Entropic Cambridge Research Laboratory, Ltd*, vol. 2, pp. 2–44, 1994.

[17] N.Obin, F. Lamare, and A. Roebel, "Syll-O-Matic : an Adaptive Time-Frequency Representation for the Automatic Segmentation of Speech into Syllables," in *International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, 2013.

[18] D. Hirst and R. Espesser, "Automatic Modelling of Fundamental Frequency using a Quadratic Spline Function," in *Travaux de l'Institut de Phonétique d'Aix*, vol. 15, 1993, pp. 71–85.

[19] P. Taylor, "The TILT Intonation Model," in *International Conference on Spoken Language Processing*, Sydney, Australia, 1998, pp. 1383–1386.

[20] C. D'Alessandro and P. Mertens, "Automatic Pitch Contour Stylization using a Model of Tonal Perception," *Computer Speech and Language*, vol. 9, no. 3, pp. 257–288, 1995.

[21] J. Teutenberg, C. Watson, and P. Riddle, "Modelling and Synthesising F0 contours with the Discrete Cosine Transform," in *International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, U.S.A, 2008, pp. 3973–3976.

[22] M. E. Beckman, J. Hirschberg, and S. Shattuck-Hufnagel, *Prosodic Typology - The Phonology of Intonation and Phrasing*. Oxford University Press, 2005, ch. The Original ToBI System and the Evolution of the ToBI Framework, pp. 9–54.

[23] A. Lacheret, J. Beliao, A. Dister, K. Gerdes, J.-P. Goldman, S. Kahane, N. Obin, P. Pietrandrea, and A. Tchobanov, "Rhapsodie : a Prosodic-Syntactic Treebank for Spoken French," in *Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland, 2014. [Online]. Available : www.projet-rhapsodie.fr

[24] J. Beliao, "Characterizing Genres through Syntax and Prosody," in *European Summer School in Logic, Language and Information*, Düsseldorf , Germany, 2013, pp. 1–12.

[25] A. Lacheret, S. Kahane, and P. Pietrandrea, *Rhapsodie : a Prosodic and Syntactic Treebank for Spoken French*. Amsterdam, Benjamins, 2015.

[26] N. Obin, "MeLos : Analysis and Modelling of Speech Prosody and Speaking Style," PhD. Thesis, IRCAM - UPMC, 2011.