AUTOMATIC MODELLING AND LABELLING OF SPEECH PROSODY: WHAT'S NEW WITH SLAM+ ?

Luigi (*Yu-Cheng*) *Liu*[123], *Anne Lacheret-Dujour*[3], *Nicolas Obin*[4]

[1]University of Ibadan, [2]IFRA Nigeria, [3]Modyco, Université Paris Nanterre & CNRS, [4]STMS, IRCAM-CNRS-Sorbonne Université

luigi.plurital@gmail.com, anne@lacheret.com, nicolas.obin@ircam.fr

ABSTRACT

This paper presents SLAM+, derived from SLAM [14], a language independent software dedicated to the data-driven melodic annotation of speech corpora, available online[1]. We discuss three main innovations introduced by SLAM: (i) the pitch format can be fixed by the user; (ii) the data quality is enhanced thanks to the integration of an additional step of pitch cleaning; (iii) the computing of two registers – global and local – enriches the acoustic processing.

Keywords: stylization, register, key and range, support and target.

1. INTRODUCTION

The prosodic annotation of speech corpora raises at least three questions: (i) the choice of a local or global model to represent and annotate the intonational characteristics of continuous speech; (ii) the methodology used for the phonetic processing of the melodic curve on which the annotation is based; (iii) the linguistic goal of the annotation.

Our work is motivated by linguistic and functional purposes. Top-down models developed in the phonological framework such as [3]; [9]; [12] and [13] were not appropriate for our research at the interface of intonation, syntax and discourse ([10]) since in these phonological approaches, the domain of projection of the pitch curve is invariably the syllable. In the *Rhapsodie* project, we needed a flexible system where the intonational domain is fixed by the user depending on the type of unit he wants to characterize: prosodic, syntactic or informational.

In contrast to a compositional phonological approach, our method is based on a global annotation of prosodic contours. While this methodology is not new ([1];[4];[8]), it has never given rise to an automatic labelling of speech prosody. Furthermore, intonational modelling can be used for two linguistic goals: typological studies designed to model the intonational contours of a language, and pragmatic analysis in order to explain how prosody is used on-line in the message. While in a typological approach, the variability of contours due to the context is disregarded, in pragmatic studies it must be taken into account. Hence, one needs to account for register dynamics, i.e. to differentiate local and global intonational registers. This point is not taken into account in the models currently used.

Section 2 presents SLAM model, a data-driven tool for the automatic modelling and labelling of speech prosody. In Section 2.1, we recap the basic characteristics of SLAM (cf. [14] for an earlier presentation). The improvement of phonetic preprocessing in SLAM that motivated the development of SLAM+ tools is presented in section 2.2: Speech cleaning (2.2.1); pitch stylization (2.2.2) and register modelling (2.2.3). Section 3 is devoted to the discussion and conclusion.

2. THE SLAM MODEL

2.1. SLAM PRINCIPLES

In the original SLAM Model, the contour of F0 (fundamental frequency) is represented by a set of three acoustic values for each unit (Fig. 1):

- Initial: the initial value of the F0 on the unit that corresponds to the first F0 value for which speech is considered as voiced.
- Final: the final value of the F0 on the unit that corresponds to the last F0 value for which speech is considered as voiced.
- Main saliency: the value corresponding to the most salient F0 peak with its time position if one exists.

Frequency values are expressed in semitones relative to the overall mean F0 of the *speaker* (Table 1). Time positions are expressed relative to the boundaries of the unit: first, middle or last part of the unit.

[1] https://github.com/vieenrose/SLAMplus

Figure 1: Acoustic representation of a melodic contour.

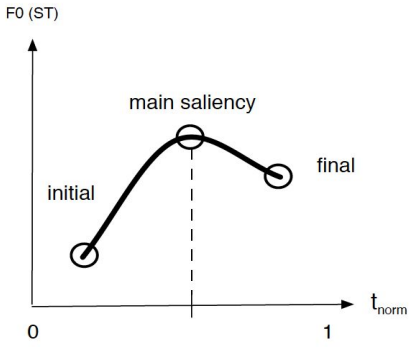


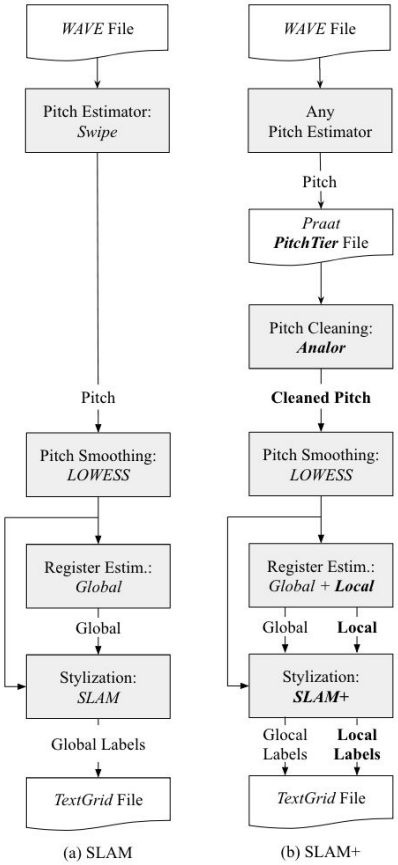
Table 1: Pitch levels used for the representation.

ELEMENTARY TONES	DESCRIPTION	RANGE (STs)
H	extreme-high	> +6
h	high	+2/+6
m	medium	-2/+2
l	low	-2/-6
L	extreme-low	< -6

2.2. PHONETIC PREPROCESSING IN SLAM+

The different steps of phonetic processing in SLAM and SLAM+, presented in the following section, are summarized in Fig. 2.

Figure 2: Processing flow of SLAM and SLAM+



The smoothing method, used to reduce the impact of signal irregularities, remains the same as in SLAM, i.e. the LOWESS algorithm [7]. It will therefore not be discussed further here.

2.2.1. Pitch cleaning

Phonetic cleaning is even more crucial nowadays as scholars mainly study ordinary speech. The quality of this type of speech is highly variable. Consequently, far from being resolved, pitch tracking errors remain frequent even in the best algorithms such as SWIPE used in SLAM ([5]). Yet, for linguistic studies, it is essential to work on a reliable melodic curve in order to develop an accurate formal and functional model of intonation. For this reason, in addition to SWIPE, a manual pitch correction step was integrated in SLAM+. Pitch cleaning is performed with Analor software ([2]) on the Praat PitchTier file of a sample. Fig. 3(a) presents a Praat pitch tier in input. In the bell curve on the left, we see the distribution of the pitch in the sample. With this tool, the user can browse the extreme values (extra-high or extra low) in the time intervals concerned and delete or modify these values if necessary, in order to obtain a clean pitch as in Figure 3(b).

Figure 3(a): Input pitch to be cleaned.

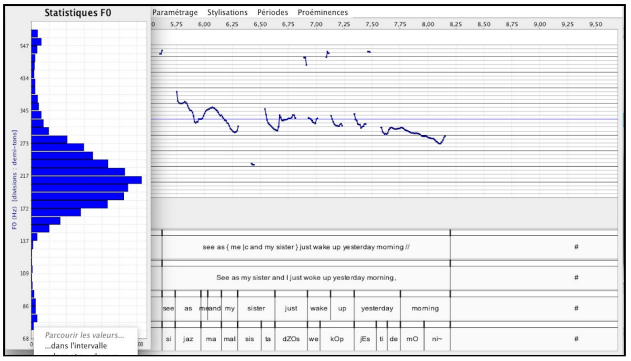
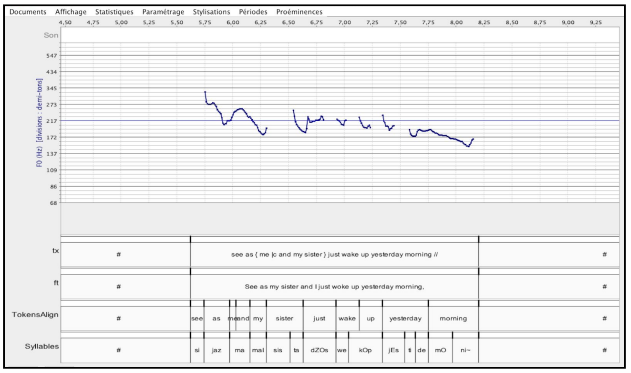


Figure 3(b): Output cleaned pitch of 2(b).



2.2.2. Pitch Stylization

- **Framework of Stylization**

The method that SLAM uses to represent the pitch contours over a user-specified linguistic unit that we call a "target" (see below), comprises three steps (Fig.1; Table 1): (i) transform the curve into coordinates relative to the reference times and frequency: initial and final times are respectively mapped to 0 and 1, while the key of the speaker register is mapped to 0 semitone; (ii) sample the frequency values of these two boundary points, sample also the maximum peak if the peak frequency exceeds endpoint frequencies from a certain threshold θ (chosen as 2 semitones); (iii) quantize the sampled frequencies in 5 regions with equal step size Δ (equal to 4 semitones), quantize the time position of the peak with a step size of 1/3 of the corresponding pitch duration. In SLAM+, the same framework of stylization is used. What is different is that the step size of frequency quantization Δ and the peak value sensitivity θ are adapted to the (dynamic) register range.

2.2.3. Account for Register Dynamism

- **"Support" vs. "Target"**

In the SLAM model, the reference frequency is set as the key of the intonational register for a speaker. This key is computed by taking the average pitch frequency over the *speaker*. In monologues, *speaker* corresponds to the audio file as a whole, while in dialogues, *speaker* corresponds to a time interval associated to one and only one speaker that is before any overlap or change of speaker. However, in the case of a monologue, it cannot be excluded that the speaker may change his/her register in a stretch of discourse. Hence, we have extended the SLAM model to adapt the frequency coordinate and its quantization to the key and the register range measured over a user-specified unit that we call "the support".

The flexibility of SLAM+ over SLAM with respect to the notion of support is illustrated in Figures (4.a-c), with the same targets being words of the utterance '*good morning my people*' (ABJ_GWA_03_M, [6]). As shown in Figure 4(a), the support - the *speaker* - is fixed in SLAM whereas in SLAM+ the support can be chosen freely. Moreover, when one wants to work at the local register level, the support is the target itself, as in Figure 4(c).

Figure 4(a) SLAM: support as *speaker*

Support	*speaker*		
Target	*good morning*	*my*	*people*

Figure 4(b) SLAM+: support as *prosodic unit*

Support	*prosodic unit*	*prosodic unit*	
Target	*good morning*	*my*	*people*

Figure 4(c) SLAM+: support as target

Support	*word*	*word*	*word*
Target	*good morning*	*my*	*people*

- **Global vs. Local Registers**

We distinguish two types of intonational registers: global and local registers. When the support of a target is chosen as the maximum extent of the register involved in this target, the register estimated over the support is considered as the global register. On the contrary, if the support is chosen as the same interval as the target's, the register estimated in this case is a local register.

- **Dynamism of Key of Register**

We propose in SLAM+ an estimation of the key of the intonational register based on the target and support specified by the user. Instead of computing the key by taking the average of F0 over the support as in SLAM, we take a weighted sum of pitch P(n) centered on the target (Fig. 5):

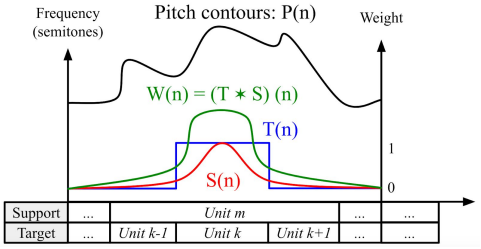
$$key = \frac{1}{C} \sum_{n \in support} P(n) \cdot W(n)$$

where C is defined as the sum of $W(n)$ over the support and $W(n)$ is the resulting window function defined as the linear convolution of a hard target function $T(n)$ and a mother window function $S(n)$ as below

$$W(n) = T(n) * S(n),\ n \in support$$

where $T(n)$ is equal to 1 if n is in the given target, and 0 if outside the target. Note that $S(n)$, chosen as *Hann* function by default, makes it possible to configure the "slope" of the boundaries of the resulting window $W(n)$ in order to take into account the temporal uncertainty of the segment boundaries for local register estimation.

Figure 5: Illustration of the proposed approach for estimation of key of range based on window method.



- **Dynamism of Register Range**

The register range corresponds to a measure which characterizes the gap between the floor pitch and the ceiling pitch. This measure is chosen such that each of these two frequency regions tolerates only α percent for saturation effect:

$$range = 2 \times max\left(|q(100-\alpha) - key|, |key - q(\alpha)|\right)$$

where *q(x)* stands for the x-th percentile of the pitch over the given support, and α is a parameter of range sensitivity. A recommended value is located between 1 and 5.

We recall that in SLAM, the frequency quantization step size Δ is set as 4 semitones for 5 regions (i.e. '*L*','*l*','*m*','*h*' and '*H*') regardless of the dynamic range of register. To better account for a broad range of registers, the frequency step size Δ is extended to

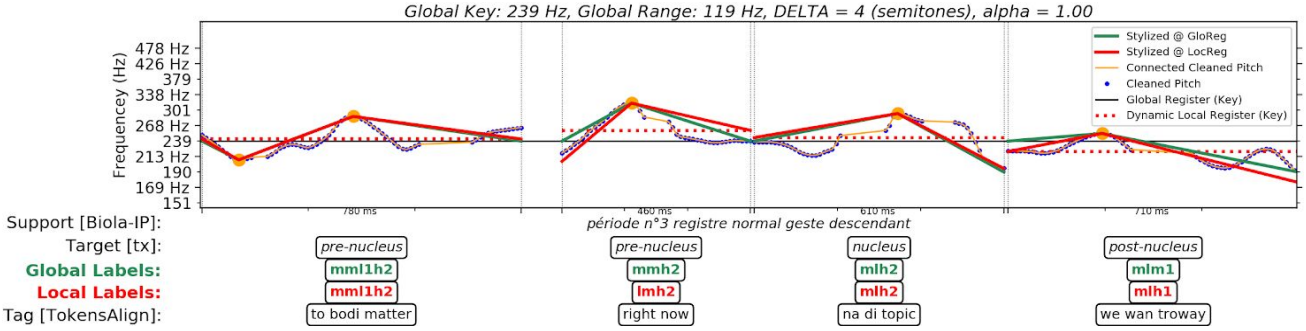
$$\Delta = max\left(\frac{range}{5}, \Delta_{min}\right)$$

where *range* stands for the estimate of the current register range, Δ_{min} is non-negative parameter which specifies the minimum step size. When $\Delta_{min} = 0$, the full dynamic range of quantization is entirely adapted to the range of register while with $\Delta_{min} = 4$, the frequency quantization is similar to the original SLAM model. It should be noted that the peak sensitivity θ, θ is set as $\Delta / 2$ in accordance with this extension.

2.3 ILLUSTRATION OF OUTPUT

Figure 6 shows the output of a SLAM+ workflow for the sample, extracted from the *NaijaSynCor* treebank (WAZA_08_M): *to bodi matter right now na di topic we wan troway* ('to body matter, right now it is the topic we want to throw away.'). The two combinations, i.e. support equal to target and support larger than target, are respectively shown in red and green lines and labels. While the local and global contours of the first prenucleus are the same, two different contours are generated for the second prenucleus (medium level at the beginning vs. low level).

Figure 6: Output of SLAM+ workflow



3. CONCLUSION AND DISCUSSION

The design of the SLAM model began 5 years ago in order to process the *Rhapsodie* resource, a syntactic and prosodic treebank of spoken French ([11]). The main specificities of SLAM can be summarized in 4 points. (i) Unlike previous methods grounded in metrical-autosegmental theories, the annotation is fully data-driven based exclusively on acoustic cues. (ii) The prosodic representation is not made on discrete fixed phonetic targets but on global melodic contours of a large set of units of various sizes and linguistic types (from the syllable to the utterance). (iii) Dialogal speech can be robustly processed even in overlapping contexts. (iv) The time component of pitch contours is considered as more trustworthy than its frequency component thanks to the systematic manual verification of the segmentation before the processing of the data. In comparison to other models of intonation, which consist in adjusting time (detection of boundaries) and frequency domain representation (estimation of key and range of register) simultaneously, the proposed approach focuses on frequency modelling.

In this paper, we have presented an extension of the original SLAM model: the SLAM+ methodology that was developed in order to provide the linguistic community with a complete data-driven package of automatic prosodic annotation. By taking register flexibility into account, this package, available online, can be used for different linguistic purposes, both typological (system-based approach) and pragmatic (usage-based approach); and, thanks to the new cleaning module, it can process any type of speech recording, from excellent to very poor. As part of our studies of the prosodic marking of informational structure, the performances of SLAM+ will be illustrated at ICPhS 2019 with the prosodic processing of topical units in the *NaijaSynCor* corpus [6].

5. ACKNOWLEDGEMENTS

This work is funded by the French National Agency for Research (ANR) through the project *NaijaSynCor* (ANR-16-CE27-0007).

4. REFERENCES

[1] Aubergé, V. 1991. *La Synthèse de la Parole: des Règles aux Lexiques*. PhD dissertation, Université Pierre Mendès-France, Grenoble, France.

[2] Avanzi, M., Lacheret, A. & Victorri, B. 2008. Analor, a Tool for Semi-automatic Annotation of French Prosodic Structure. In Proceedings of Speech Prosody, 119-122. Campinas, Brazil.

[3] Beckman, M. E., Hirschman, J., Shattuck-Hufnagel, S. 2005. The original ToBI system and the evolution of the ToBI framework. In: Jun, S-A. (ed), *Prosody typology : The phonology of intonation and phrasing*, Oxford: Oxford University Press, 9-54.

[4] Bolinger, D. 1989. *Intonation and its uses,* London: Arnold.

[5] Camacho, A. 2007. *SWIPE: A sawtooth waveform inspired pitch estimator for speech and music*. Gainesville: University of Florida.

[6] Caron, Bernard. 2017. NaijaSynCor : A corpus-based macro-syntactic study of Naija (Nigerian Pidgin). http://naijasyncor.huma-num.fr/ (10 December, 2018).

[7] Cleveland, W. S. 1981. LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *American Statistician*, *35*(1), 54.

[8] Delattre, P. 1966. Les Dix Intonations de Base du Français. *The French Review*, 40(1), 1-14.

[9] Delais-Roussarie, E., Post, B., Avanzi, M., Buthke, C., Di Cristo, A., Feldhausen, I., Sun-Ah, J., Martin, P., Meisenburg, T., Rialland, A., Sichel-Bazin, R. & Yoo, H.-Y. 2015. Intonational phonology of French: developing a ToBI system for French. In: Sonia Frota & Pilar Prieto (eds), *Intonation in Romance*, Oxford: Oxford, Linguistics, 63-100.

[10] Lacheret, A., Kahane, S., Beliao, J., Dister, A., Gerdes, K., Goldman, J. P., Tchobanov, A. 2014. Rhapsodie: a prosodic-syntactic treebank for spoken French. *Proc*. *LREC* Reykavik, Lrec-conf.org

[11] Lacheret-Dujour, A., Kahane, S. Pietrandrea, P. 2019. Rhapsodie*: a prosodic and syntactic treebank of spoken French*, Amsterdam: Benjamins.

[12] Looze, C. D., Hirst, D. 2010. Integrating changes of register into automatic intonation analysis. *Proc Speech Prosody,* Chicago.

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

[14] Obin, N., Beliao, J., Veaux, C., & Lacheret, A. 2014. SLAM: Automatic stylization and labelling of speech melody. *Proc Speech Prosody*, Dublin, 246-250.