At the Interface of Speech and Music: A Study of Prosody and Musical Prosody in Rap Music

Olivier Migliore[1], *Nicolas Obin*[2]

[1] RIRRA 21, Université Paul Valéry - Montpellier III, Montpellier, France
[2] IRCAM, CNRS, Sorbonne Université, Paris, France

Abstract

This paper presents a pioneer study of speech prosody and musical prosody in modern popular music, with a specific attention to music where the voice is closer to speech than to sing. The voice in music is a complex system in which linguistic and musical systems are coupled and interact dynamically. This paper establishes a new definition of the musical prosody in order to model the specific relations between the voice and the music systems in this kind of music. Additionally, it presents a methodology to measure the musical prosody from the speech and music signals. An illustration is presented to assess whether the speech prosody and the musical prosody can characterize the phonostyle of a speaker, by comparison of three American-English rappers dating from the beginning of the 2000's. The main finding is that not only the rappers can be characterized and distinguished by their speech prosody, but also by their musical prosody, i.e. by the degree of synchronization between their lyrics with the musical system.

Index Terms: speech and music, speech prosody, musical prosody, popular music

1. Introduction

Though speech prosody and its application to the study of stylistics is now well-established in the speech community [1, 2, 3], its extension to the musical domain remains rare and limited. The reasons are manifold: from the complexity of the voice/music system, the diversity of the voice in music, and the limited resources available. First, the voice/music system is a complex system in which the linguistics and the musical systems interact dynamically. Second, the voice has a large spectrum in the history of music from spoken to singing voice: classical singing attracting most of the attention of musicological and linguistic research. Finally, the study of the voice in the 20th century popular music must also face the lack of resources: the sound recording is the only resource available [4] and the voice is usually mixed with the musical background. Accordingly, most of studies on the voice in popular music do not study the voice itself but are limited to the external traces available, such as lyrics [5, 6], sociological [7, 8] and phonological(accentuation from text, see [9, 10, 11] on folk French and English songs) aspects. In particular, the *musical prosody* has been introduced to study the relationship of voice and music [12]. This however remains highly limited due to the derivation of the supposed accentuation from the texts (linguistic and musical), but not from the sound signal. Finally, recent studies have indicated the importance of considering the sound signal to study the voice in order to study the interpretative nuances, the stylistics, and the complex relationship between voices and music in popular music.[13, 14, 15].

This paper investigates the modeling of prosody and musical prosody in popular music in which the voice is closer to speech than to sing. It addresses the two main issues: 1) how to measure the relationships of speech and music ? 2) what are the elements of speech prosody and musical prosody that are characteristics of the style of a speaker ? To answer these issues, this paper establishes a theoretical and methodological framework to define and measure the speech prosody and the musical prosody from sound recordings. Besides, the paper introduces an algorithm for the representation and visualization of prosodic contours, based on the joint clustering of the pitch contours and their corresponding duration. The proposed contributions are illustrated by a study of three rap American songs dating from the beginning of the 2000's.

2. The Musical Prosody

2.1. What, why, and how ?

Musical prosody has been a subject of important interest for poets and musicians since the ancient times, defining how the singing voice should be placed in accordance with the musical accompaniment, its metric and its harmony. In modern times, the first known study of popular music [12] defined the musical prosody as "*the concordance between musical parameters (accents, duration, intervals) and accents of the text*" [1], arguing that the musical prosody is essentially rhythmical in popular music. Though this study represents the first attempt

1. The musical prosody concerns the articulation of the voice and the music, and should not be confused with the prosody of the singing voice.

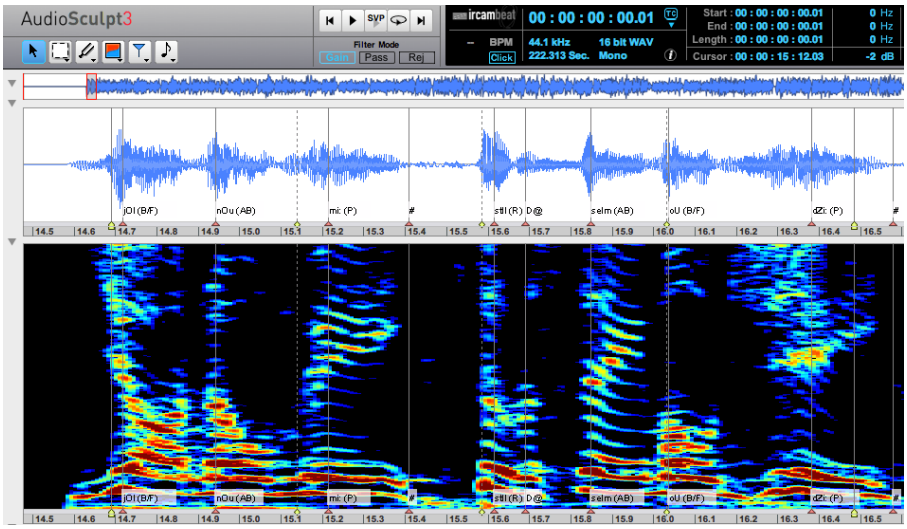

FIGURE 1 – Illustration of the segmentation and labeling on AudioSculpt for the first sentence of the songs "*Forgot about Dre*": "*You know me, still the same OG*". On top, the speech waveform, on bottom, the corresponding magnitude spectrogram. Musical beats are represented by plain and dashed vertical lines for downbeats and beats, and yellow markers. Syllables are represented by plain vertical lines and red markers placed at the position of its vowel onset, with corresponding phonetic transcription, prosody (P, R, F) and musical prosody (B, AB) labels.

to define the musical prosody, the proposed definition is however based on a number of debatable assumptions. First, the notion of concordance presupposes that there is an obligatory match between music and voice accents, which is based on normative linguistic and occidental music repertory theories assuming that accents are obligatory and exclusively dictated by the texts (lyrics and musical score) [16, 9, 10, 11]. This no longer applies in modern popular musics in which the sound recording is the only resource available, and for which the vocal and music accents are extremely free. In other words, the relation between voice and music is not fixed a priori but is constructed a posteriori in a complex and dynamic manner, shaping musical styles. Consequently, the study of the musical prosody is concerned by the description of the degree of synchronization between voice and music, which must be deducted from the analysis of the sound signal.

We propose an alternative and more general definition of the musical prosody as: "*the ratios of rhythm, quantity and accentuation between the syllables of the words and the beats of the measure*" [14]. Accordingly, analyzing musical prosody consists in examining the articulation between the rhythmic unities of linguistic and musical systems based on the actual sound realization, without prior application of musical or linguistic knowledge. In this complex voice/music system, the rhythmic placement of the stressed and unstressed syllables can be realized in concordance with the musical metric, without concordance, with a large spectrum of possibilities in between. In the remaining of this section, we describe the processing chain used to segment and annotate the rhythmic units of speech and music from the sound signals, and the parameters proposed to describe the musical prosody: the degree of synchronization between speech syllables and the musical metric frame.

2.2. Methodology

This section presents the methodology and the tools used for the segmentation and the annotation of the speech and musical rhythmic units from the music recording. This processing requires the separated speech and music tracks, and can be assisted by a computer. Annotations and visualization were processed with the AudioSculpt software [17].

2.2.1. Musical processing

The rhythmic units of the music are the musical beats and their eight notes, which are estimated from the musical mix signal by using the Ircambeat system [18]. This system estimates automatically from the music signal the global and the local time-varying tempo. It also estimates the positions of the beats and the seconds quavers of beat, further referred to as *afterbeat*.

2.2.2. Speech processing and labeling

The syllable and its vowel onset are chosen as the rhythmic units of speech prosody. In particular, the vowel onset is here considered as a fair approximation of the perceptual center of speech [19]. The segmentation of speech into syllables and vowel onset was processed automatically by using the Syll-O-Matic system [20], manually corrected and complemented with the corresponding phonetic transcriptions in the SAMPA alphabet. The description of the syllables and corresponding vowel onsets were then augmented with manual labeling of prosody and musical prosody based on the perception of an expert annotator. First, the following prosodic events were considered and labeled from the speech signal: the final accent of a prosodic phrase which fall on the last syllable of each breath group (marked as "P"), the final accent of a rhythmic group which fall on the last syllable

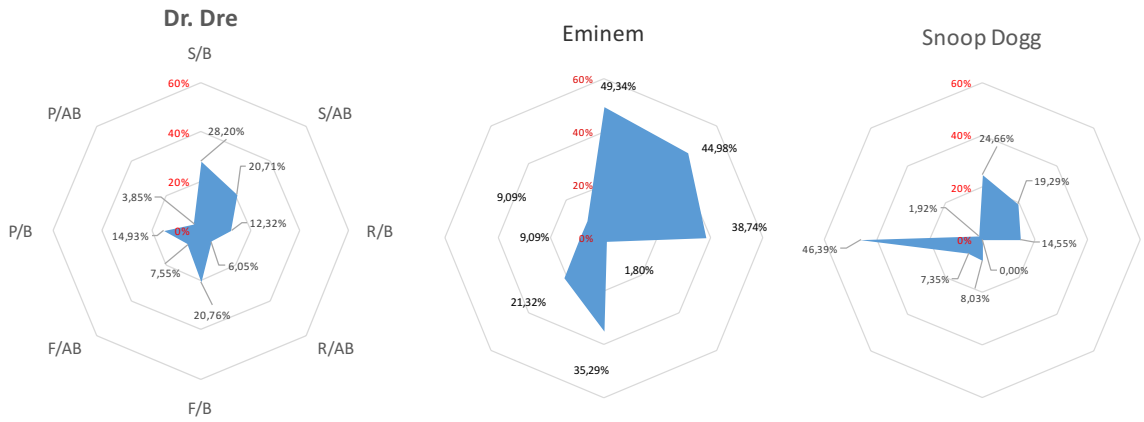

FIGURE 2 – Prosodic octagons obtained for the three rappers. From left to right: Dr. Dre, Eminem, and Snoop Dogg.

of a word within a breath group (marked as "R"), and the focus accent which concern any other than the final syllables of a word or a breath group (marked as "F"). Second, the synchronization of the syllables with the musical measure were reported from the mix signal, the vowel nuclei markers and the musical beats markers. Each vowel nuclei perceived by the annotator as falling on a musical beat is marked as "B" and on an afterbeat is marked as "AB".

2.3. Description of the Musical Prosody

The proposed processing chain can then be used to describe the musical prosody of a music track, by measuring *the degree of synchronization between speech syllables and musical beats*. This synchronization is represented by mean of a prosodic octagon, reporting eight proportions of synchronization expressed in percentage:

— Proportion of syllables which fall on beat (S/B) and on afterbeat (S/AB);

— Proportion of phrase accents which fall on beat (P/B) and on afterbeat (P/AB);

— Proportion of rhythmic accents which fall on beat (R/B) and on afterbeat (R/AB);

— Proportion of focus accents which fall on beat (F/B) and on afterbeat (F/AB).

Please note that beats and afterbeats are both investigated for synchronization in the proposed prosodic octagons. This is due to the fact that speech and music synchronization may not necessarily occur on a beat, but also on any of its subdivisions, as for instance the simplest one: the afterbeat.

3. Illustration

The proposed contribution is illustrated by a phonostylistics study of English-American hip-hop, without loss of generality to other popular musics. This illustration assesses the respective contributions of speech prosody and musical prosody in the construction of the phonostyle of a rapper, his vocal "flow" [21].

3.1. Material

The material used for this study is based on three American-English rap songs by the famous rapper Dr. Dre, all from his album *2001*: "*Still Dre*" (4'30"), "*The Next Episode*" (2'41"), and "*Forgot about Dre*" (3'42"). The audio material is composed of the acapella signal and the original mix, synchronized. The speech and music processing has been conducted as described previously in Section 2, accompanied by the estimation of the fundamental frequency of the speaker from the acapella signal by using the SWIPE algorithm [22] [2].

3.2. Musical Prosody

The first part of this study is concerned with the characterization and the comparison of the musical prosody of the different rappers. Figure 2 presents the prosodic octagons obtained for the three rappers. Globally, many stressed syllables do not fall on the beats or the afterbeats: 22.23% only falls on the beat, 6.55% on the afterbeats, and 71.22% elsewhere. This is clear evidence for the fact that the synchronization between the speech and the music is not obligatory, contradicting prior works on musical prosody [12]. This may be even more true for rap music, for which the deviation to the linguistic and musical codes testify symbolically of the "attitude" of the rapper towards social codes [14]. Besides, the octagons obtained for the rappers shows strongly different patterns, highlighting their vocal specificities. Once again, DD is globally the most classical rapper since he synchronizes most of his stressed syllables on beats with a relatively small variations across them (P/B=14.93%, R/B=12.32%, F/B=20.76%), as compared to the other rappers. Besides, SD is the less (S/B=24.66%, S/AB=19.29%) and E is the most (S/B=49.34%, S/AB=44.98%) synchronized with the musical measure. On the one side, E is highly synchronized especially with rhythmic and focus accents on beats (R/B=38.74% and F/B=35.29%) and also on afterbeats (S/AB=44.98%), but surprisingly places his

2. The analyses created for this study are made freely available for research at: www.github.com/nicolasobin/

phrase accents out of the beat (P/B=9.09%). On the other side, SD has the most heterogeneous behavior: it largely synchronizes his phrase accent to the musical measure (P/B=46.39%) to the detriment of the other accents (R/B=14.55%, F/B=8.03%). This clearly shows that rappers also use their musical prosody to construct their *vocal style*.

3.3. Speech Prosody

The second part of this study is concerned with the characterization and the comparison of the prosodic contours of the different rappers. To do so, a clustering of the prosodic contours of the rappers is conducted so as to reveal and compare their main prototypes. Contrary to the existing prosodic clustering techniques [23, 24, 25] which are focused on the pitch contour only, this paper establishes a simple algorithm for the clustering of pitch and duration contours by mean of a weighted k-means algorithm [26, 27].

Formally, let $\mathbf{x} = [x_1, \ldots, x_N]$ a vector describing a data point. The objective function of the proposed weighted k-means clustering is defined as:

$$W(\mathcal{C}; K) = \sum_{i=1}^{K} \sum_{x \in \mathcal{C}_i} \sum_{j=1}^{N} w_j ||x_j - \mathcal{C}_j^i||^2 \qquad (1)$$

where K is the number of clusters, $\mathcal{C}^i$ is the centroid of the i-th cluster, $\mathbf{x} \in \mathcal{C}^i$ denotes that $\mathcal{C}^i$ is the closest centroid to $\mathbf{x}$, and w_j is the weight of the j-th element of $\mathbf{x}$. The clustering is obtained by minimizing this objective function analogously to the classical k-means algorithm. In the present study, we define $\mathbf{x}$ as $[\mathbf{f_0}, d]$, where $\mathbf{f_0}$ is the time-normalized vector of pitch values forming the pitch contour (in Hz), and d the duration of the pitch contour (in ms). Here, the pitch contour is estimated on each syllable as the longest sequence of voiced pitch values, and resampled on $N_{f0} = 50$ values to construct a time-normalized pitch contour vector. Thus, the weight w_j is set in order to balance the importance of the pitch contour vector and the duration during clustering. This is simply done by fixing w_j to the inverse of the dimension of the corresponding vector, i.e., $w_j = 1/N_{f0}$ for the pitch values, and $w_j = 1$ for the duration value. The clustering has been computed on the pitch contours and duration of all syllables (including stressed and non-stressed syllables).

Figure 3 illustrates the five main prosodic contours (pitch and duration) obtained for the three rappers: Dr Dre (DD, 1,025 syllables), Eminem (E, 445 syllables), and Snoop Dogg (SD, 342 syllables). This figure exhibits that the three rappers under investigation have clear characteristics and distinctive prosodic contours, even regardless to the individual difference in range and dynamics. On the one side, some of these contours are relatively close to what could be expected in speech [28]. For instance, DD and E have variations around the classic "bell" contour

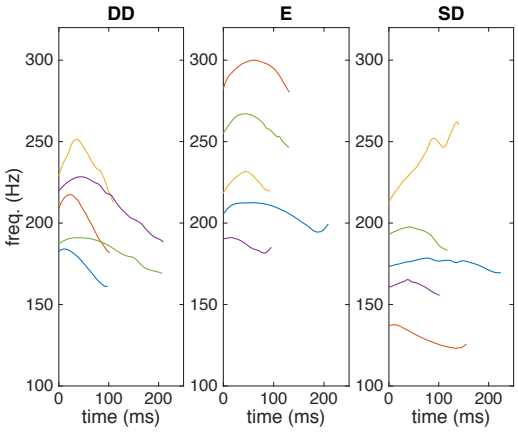

FIGURE 3 – The five main pitch/duration contours obtained for the three individuals after weighted k-means clustering: Dr Dre (DD), Eminem (E), and Snoop Dogg (SD).

widely observed in speech. DD has the most classical contours, with the specificity of being highly asymmetric with an early peak followed by a long fall, typical of his "slur" flow. Conversely, E has a large dynamics and nearly symmetric contours, with a middle peak and short durations typical of his "metronomical" flow. On the other side, some contours are clearly more unexpected, showing the freedom of rappers towards speech standards. SD is typical of this freedom: the sustained rising contour in the high pitch range which he uses for interjections and punctuations and the long contours are typical of his "unexpected" flow. Also, the patterns have important difference in duration: some short (<100ms) and some long (around 200 ms), opening a new dimension for the interpretation of prosodic contours. This clearly confirms that speech prosody is fully part of the vocal identity and vocal style of rappers.

4. Conclusion

This paper presented a stylistic study of prosody in modern popular music. To do so, the paper established a definition of the musical prosody, a methodology to measure it from the sound signal, and proposed a simple algorithm to visualize prosodic contours based on weighted clustering. The proposed prosody and musical prosody representations were illustrated to study and compare the phonostyles of three American-English rappers dating from the beginning of the 2000's. This study proved evidence that rappers used speech prosody and musical prosody to construct their vocal style. Further research will focus on investigating other possible musical prosody parameters in order to refine the description of the interpretative nuances of voices in music. Finally, the proposed methodology applies to a large variety of languages and popular musics, opening large possibilities for the study of the voice and its stylistics in popular musics.

5. References

[1] P. Léon, *Précis de phonostylistique. Parole et expressivité*. Paris: Nathan, 1993.

[2] A.-C. Simon, A. Auchlin, M. Avanzi, and J.-P. Goldman, "Les phonostyles: une description prosodique des styles de parole en français," in *Les voix des Français : en parlant, en écrivant*. Bern: Lang, 2010, pp. 71–88.

[3] N. Obin, "MeLos: Analysis and Modelling of Speech Prosody and Speaking Style," PhD. Thesis, Ircam - Upmc, 2011.

[4] O. Julien, "L'analyse des musiques populaires enregistrées," *Observatoire musical français*, no. 37, pp. 141–166, 2008.

[5] B. Ghio, "Littérature populaire et urgence littéraire : le cas du rap français," *TRANS-*, no. 9, 2010.

[6] D. Rossi, "Le vers dans le rap français," *Cahiers du Centre d'études métriques*, no. 6, pp. 115–143, 2012.

[7] K. Hammou, *Une histoire du rap en France*. Paris: La découverte, 2012.

[8] A. Mehrabian, *Voix du rap. Essai de sociologie de l'action musicale*. New-York: L'Harmattan, 2007.

[9] C. Palmer and M. Kelly, "Linguistic prosody and musical meter in song," *Journal of Memory and Language*, vol. 31, p. 525–542, 1992.

[10] F. Dell and J. Halle, "Comparing musical textsetting in French and in English songs," in *Towards a Typology of Poetic Forms*, J.-L. Aroui and A. Arleo, Eds. Amsterdam: John Benjamins, 2009, pp. 63–78.

[11] N. Temperley and D. Temperley, "Stress-meter alignment in French vocal music," *Journal of the Acoustic Society of America*, vol. 134, no. 1, pp. 520–527, 2013.

[12] B. Joubrel, "Approche des principaux procédés prosodiques dans la chanson francophone," *Musurgia*, vol. 9, pp. 59–70, 2002.

[13] C. Chabot-Canet, "Interprétation, phrasé et rhétorique vocale dans la chanson française depuis 1950 : expliciter l'indicible de la voix," PhD. Thesis, Université Lyon II-Louis Lumiere, Lyon, France, 2013.

[14] O. Migliore, "Analyser la prosodie musicale du punk, du rap et du ragga français (1977-1992) à l'aide de l'outil informatique," PhD. Thesis, Université Paul-Valéry Montpellier 3, 2016.

[15] M. Ohriner, "Metric ambiguity and flow in rap music: A corpus-assisted study of outkast's "mainstream" (1996)," *Empirical Musical Review*, vol. 11, no. 2, pp. 153–179, 2016.

[16] M. Gribensky, "Prosodie et poésie. place des études sur la prosodie poético-musicale dans la recherche musico-littéraire (bilan et perspectives)," *Fabula / Les colloques*, 2010.

[17] N. Bogaards and A. Roebel, "An interface for analysis-driven sound processing," in *Convention of the Audio Engineering Society*, 2005.

[18] G. Peeters and H. Padapopoulos, "Simultaneous beat and downbeat-tracking using a probabilistic framework: theory and large-scale evaluation," *IEEE Xplore Digital Library*, vol. 19, no. 6, 2011.

[19] C. A. Fowler, ""perceptual centers" in speech production and perception," *Perception & Psychophysics*, vol. 25, no. 5, pp. 375–388, 1979.

[20] 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.

[21] K. Adams, "On the metrical techniques of flow in rap music," *Society for Music Theory*, vol. 11, no. 5, 2009.

[22] A. Camacho, "SWIPE: A Sawtooth Waveform Inspired Pitch Estimator for Speech and Music," PhD. Thesis, University of Florida, 2007.

[23] U. D. Reichel, "Data-driven extraction of intonation contour classes," in *ISCA Workshop on Speech Synthesis*, 2007, pp. 240–245.

[24] M. Gubian, F. Cangemi, and L. Boves, "Automatic and Data Driven Pitch Contour Manipulation with Functional Data Analysis," in *Speech Prosody*, 2010, pp. 181–189.

[25] D. Sacha, Y. Asano, C. Rohrdantz, F. Hamborg, D. Keim, B. Braun, and M. Butt, "Self Organizing Maps for the Visual Analysis of Pitch Contours," in *Nordic Conference of Computational Linguistics*, 2015, pp. 181–189.

[26] G. Tseng, "Penalized and weighted k-means for clustering with scattered objects and prior information in high-throughput biological data," *Bioinformatics*, vol. 23, no. 17, p. 2247–2255, 2007.

[27] M. Ackerman, S. Ben-David, S. Branzei, and D. Loker, "Weighting clustering," in *Association for the advancement of artificial intelligence*, 2012, pp. 858–863.

[28] N. Obin, J. Beliao, C. Veaux, and A. Lacheret, "SLAM: Automatic Stylization and Labelling of Speech Melody," in *Speech Prosody*, 2014. [Online]. Available: https://github.com/jbeliao/SLAM/