Novel Metrics of Speech Rhythm for the Assessment of Emotion

Fabien Ringeval, Mohamed Chetouani, Björn Schuller

Novel Metrics of Speech Rhythm for the Assessment of Emotion

Fabien Ringeval[1,3]*, Mohamed Chetouani*[2]*, Björn Schuller*[3]

[1]DIVA group, Department of Informatics, University of Fribourg, Switzerland
[2]Institut des Systèmes Intelligents et de Robotique, Université Pierre et Marie Curie, Paris, France
[3]Institute for Human-Machine Communication, Technische Universität München, Germany

fabien.ringeval@unifr.ch, mohamed.chetouani@upmc.fr, schuller@tum.de

Abstract

Whereas rhythmic speech analysis is known to bear great potential for the recognition of emotion, it is often omitted or reduced to the speaking rate or segmental durations. An obvious explanation is that the characterisation of speech rhythm is not an easy task itself and there exist many types of rhythmic information. In this paper, we study advanced methods to define novel metrics of speech rhythm. Their ability to characterise spontaneous emotions is demonstrated on the recent Audio/Visual Emotion Challenge Task on 3.6 hours of natural human affective conversational speech. Emotion is assessed for the four dimensions Activation, Expectation, Power and Valence as binary classification tasks on the word level. We compare our new rhythmic feature types to the official 2 k bruteforce acoustic baseline feature set on the Audio Sub-Challenge. In the results, the rhythmic features achieve a promising relative improvement of 16% for Valence, whereas the performance is more contrasted for the three others dimensions.

Index Terms: speech rhythm, prosodic features, emotion recognition

1. Introduction

Most definitions from the literature consider rhythm as being conveyed by the perceived information during the alternation (or repetition) of events spaced over time. However, this definition considers events from various origin such as: (i) biological (e.g., food, heart, respiratory, etc.), (ii) body (e.g., choreography) or (iii) from speech [1]. Rhythm thus refers to the notion of dynamic movement in the perception of different types of phenomena. Because these phenomena are very diverse, their entanglement is evident for speech [2], and since the mechanisms of human perception are also very complex [5], identifying in concrete and simple terms the characteristics of rhythm is particularly difficult.

This paper investigates different metrics of speech rhythm, with the aim to study their relevance for the characterisation of spontaneous emotions from natural human conversational speech. The goal of this work is to help provide new relevant features for the analysis of affective spontaneous interactions. In the remainder of this paper we briefly introduce some rhythmic phenomena that were identified in the literature (Sec. 2), the metrics of speech rhythm (Sec. 3), including the taxonomical models (Sec. 3.1) and the new advanced ones (Sec. 3.2), the experimental setup (Sec. 4), including the database (Sec. 4.1) and the rhythmic feature set (Sec. 4.2) and present experimental results (Sec. 5) before concluding (Sec. 6).

2. Speech Rhythm

A literature survey on rhythm shows how difficult it is to have a precise definition of what it refers to, since many conceptual and terminological inventories are available [3]. However, a set of pretty well established phenomena have been identified so far, such as: (i) the duality between form and structure [4], (ii) some temporal distortions [5] and (iii) some preferential anchors according to the language [1].

The majority of studies that were conducted on speech rhythm were guided by a taxonomical spirit, i.e., in order to perform a classification of languages [1]. The emergence of new metrics of rhythm in the last decade brought a revival of interest in the taxonomical community. Temporal properties of consonants and vowels were used to argue for the existence of a rhythmic continuum between stress (e.g., English and German) and syllabic (e.g., Spanish and French) languages [6].

Because emotions clearly rely on dynamic processes in both their production and perception counterparts [7], new advanced models of speech rhythm would help to provide relevant features for their characterisation. Indeed, the automatic processing of both spontaneous and natural emotions still remains a challenge, specifically when it comes to real data analysis.

3. Metrics of Speech Rhythm

We describe in the following paragraphs several techniques that have been proposed to characterize the rhythmic information conveyed by speech. First we present the techniques used by the taxonomical community and then we describe our novel metrics of speech rhythm.

3.1. Taxonomical Models

The taxonomical models of speech rhythm use measurements of the segmental duration or the quantity of segments according to a given speech unit (e.g., vowels, consonants, words, etc.). The speech rate is, for example, often used as a unique feature of rhythm in emotion recognition systems [8], although many studies have shown that it is only one component [9]. In addition, there are many metrics whose use for the characterisation of affective correlates from speech could be studied.

3.1.1. Vocalic and Consonantal Variability Phenomena.

Ramus et al., proposed a measure of speech rhythm based on the percentage of vocalic intervals (%V) and the standard deviation of consonantal intervals (ΔC), with the aim of quantifying a rhythmic continuum between stress and syllabic languages [10]. However, these measures would only be relevant for the study of a corpus for which the speech rate is strictly controlled.

3.1.2. Compensatory Phenomena (Oscillatory Mechanisms).

Brady et al., used circular statistical measures to study the cognitive processes of speech for Japanese [11]. After a detection of the attacks of voiced syllables, a sinusoidal waveform was generated with a period set by the average interval duration between the segments. Therefore, the position of segments in time is represented by a value of phase θ_i in the generated sinusoid. The periodicity of segments, $\bar{R}$, was then quantified as the sum of the vectors corresponding to the values of phase θ_i, divided by the number *N* of segments.

$$\bar{R} = \frac{1}{N}\left(\left(\sum_{i=1}^{N}\sin 2\pi\theta_i\right)^2 + \left(\sum_{i=1}^{N}\cos 2\pi\theta_i\right)^2\right)^{1/2} \quad (1)$$

3.1.3. Variation Coefficient of Segmental Duration.

The coefficient of variation ($Varco$), which is defined as the ratio between the standard deviation σ and the mean μ of a given distribution, was used on vocalic and consonantal intervals by [12]. This measure was combined with that of %V and allowed to discriminate syllabic from stress languages. However, the differences were found to be at least as significant between the dialects of these languages.

3.1.4. Pair-wise Variability Index.

Grabe and Low proposed to measure the temporal variability of pairs of successive phonetic intervals (I_k and I_{k+1}) by using the *rPVI* [13] (2); a normalisation to the speaking rate has been proposed with the *nPVI* (3). These measures helped to strengthen the theory of rhythmic language classes outlined above.

$$rPVI(k) = I_{k+1} - I_k \quad (2)$$

$$nPVI(k) = 2\frac{I_{k+1} - I_k}{I_k + I_{k+1}} \quad (3)$$

Other studies suggested that the comparison of time intervals between phonetic units could be achieved by using their ratio, instead of their difference [6]. The rhythm ratio (*RR*) measure (4) provided results close to the *nPVI* on corpora of languages.

$$RR(k) = \frac{I_k}{I_{k+1}} \quad (4)$$

3.2. Advanced Models

Some authors have proposed to expand the definitions of the metrics used in the taxonomical models [14]. It was suggested that the phenomena of rhythm could be generated by the dynamics in prosody. Indeed, Lerdahl et al. showed that pitch serves to distinguish the accents of music in groups of: (i) metric, (ii) phenomena and (iii) structures [15]. Moreover, prosodic particularities seem to be related to the strong beats of rhythm, since they fall at: (i) important changes or low values in the pitch, (ii) changes in the harmonics or (iii) changes in rates.

In the following, we first present two advanced methods of features extraction that use the repartition of speech units over time to quantify rhythm. Whereas the final two techniques use both temporal information and changes in the prosodic shape of consecutive speech units to characterise speech rhythm.

3.2.1. Low Fourier Frequency Analysis.

Tilsen et al. proposed a method to extract the rhythmic envelope of the speech signal [16]. A low frequency (LF) signal is calculated on the speech signal by several filtering steps, which are supposed to represent the process of perception of rhythm. As the waveform of the LF signal is stationary, a Fourier transform can be used to estimate the values of entropy, centroid and the average frequency from the rhythmic envelope.

3.2.2. Instantaneous Frequency and Envelope.

We proposed in [17] to use the Hilbert-Huang Transform (HHT) for characterising speech rhythm. A speech unit interval (SUI) signal is first generated by a resampling process (cubic spline, Fs = 32 Hz) on interval durations of speech units. Because the interval duration between phonemes is known to vary from 60 ms to 1 second (i.e., from 1 Hz to 16 Hz)[18], all frequencies of speech rhythm can be captured with a sampling frequency set to 32 Hz. Empirical mode decomposition (EMD) is then applied on the SUI signal to extract the HHT derived features: instantaneous amplitude and frequency from the sum of the three first intrinsic mode functions (provided by the EMD), and the mean instantaneous frequency (MIF) obtained by the calculation proposed in [19].

3.2.3. Prosodic Pair-wise Variability Index.

In order to characterize the rhythm through the dynamic of prosody, we propose to extend the *PVI* [13]. The time interval measurement I_k is replaced by the *Varco* coefficient of given prosodic low-level-descriptor (LLD), such as pitch, loudness, spectral flux, ... [20]. A normalisation factor α is used to take into account durations d of consecutive segments k and $k+1$ and their interval duration I_k in the *pPVI* metric:

$$pPVI(k) = \alpha.\left(Varco_{k+1} - Varco_k\right) \quad (5)$$

$$\text{with } \alpha = \log\left(\frac{d_k d_{k+1} I_k}{d_k + d_{k+1} + I_k}\right) \text{ and } Varco = \frac{\sigma}{\mu}$$

The value of this feature is equal to zero if the dispersion of the LLD is identical on two consecutive segments of speech, which means a monotony in the prosodic component. Otherwise, the values depend on the amount of changes present in the *Varco* of the LLD between the speech segments. Due to the normalisation factor α, the values of *pPVI* also depend on both the duration and interval of the two consecutive segments; these effects are cumulative. The logarithm of the duration ratio was computed to reduce its variability. Finally, a local maximum (or minimum) in the *pPVI* defines prominence in a given prosodic LLD.

3.2.4. Prosodic Hotelling Distance.

The Hotelling distance (HD) is a measure for comparing the statistical distribution of two data sets by a calculation similar to the Mahalanobis distance. In particular, it involves a normalisation factor by the duration of the two compared speech segments. However, as the interval duration between these segments is not included in the HD calculation, we added this value using the normalisation coefficient α. This new metric is termed the prosodic Hotelling distance (*PHD*):

$$PHD(k) = \alpha\left[\left(\mu_k - \mu_{k+1}\right)^T \Sigma^{-1}_{k\cup k+1}\left(\mu_k - \mu_{k+1}\right)\right] \quad (6)$$

where μ_k, μ_{k+1} denote the means of a prosodic LLD on the past (k) and new ($k+1$) speech segments, and $\Sigma_{k\cup k+1}$ the covariance matrix of the joint events k and $k+1$.

This measure can be used for one or several LLDs, and two different techniques are available to define the matrix $\Sigma_{k\cup k+1}$:

Table 1: Overview of the AVEC dataset per partition

	Train	**Develop**	**Test**
# Sessions	31	32	32
# Words	20 183	16 311	13 856
Avg. word dur. [ms]	262	276	249

Table 2: Overview of class balance: fraction of positive instances over total instances of words in the training and developing partitions.

Ratio	ACT.	EXP.	POWER	VALENCE
Train	0.496	0.409	0.560	0.554
Develop	0.581	0.334	0.670	0.654

(i) the first consists of filling the diagonal with the standard-deviation values of each LLD and (ii) the second technique exploits all values from the covariance matrix.

The value of the *PHD* is equal to zero when the distributions of the prosodic LLD(s) are identical between pairs of consecutive segments, and positive in all other cases. It varies proportionately with the amount of change present in the statistical distribution of the LLD(s), and the normalisation factor α also influences the values of the *PHD*, like for the *pPVI*.

4. Experimental Setup

In this section we describe the methodology we used for the AVEC emotion recognition Audio Sub-Challenge. Our approach consisted of comparing the relevance of our rhythmic features with the official prosodic 2 k brute-forced acoustic baseline feature set.

4.1. Database

The Solid-Sensitive Artificial Listener (SAL) part of the SEMAINE corpus was used for the AVEC challenge [21]. In this database, participants were asked to talk in turn to four emotionally stereotyped human operators. The used language was English and all the sessions were split in three: a training, develop, and test partition. Table 1 shows the distribution of data in sessions and words for each partition. Activation, Expectation, Power, and Valence compose the annotated emotion dimensions of the Solid-SAL corpus. The binary labels of each affective dimension were obtained at the word level by a threshold on the continuous values that were rated using the tool called Feeltrace. Table 2 lists the fraction of positive instances per partition and per dimension from the Solid-SAL corpus.

4.2. Rhythmic Features

The taxonomical models of speech rhythm provide 6 LLDs at the word level: duration, interval, *rPVI*, *nPVI* and *RR* (for both duration and interval). As functionals are computed on these LLDs for each word, values were resampled by cubic splines (Fs = 32 Hz). 30 functionals (cf. Table 1 in [17] plus second order regression coefficient, mean and standard-deviation value of raising / falling values and mean number of raising / falling values) were then computed on the 6 LLDs. 4 additional features: word rate, periodicity, and *Varco* of durations and intervals were merged, so that 184 features of speech rhythm were returned in total by the taxonomical models.

The new advanced models provide much more features. The *pPVI* and *PHD* metrics were performed at the word level on each LLD returned by the openSMILE feature extraction's toolkit [22]; 25 energy and spectral related LLDs, plus 6 voicing related LLDs with delta coefficients. Obtained values were resampled by cubic splines (Fs = 32 Hz) and the 30 functionals were computed at the word level. Finally, the 30 functionals were also applied on the HHT derived features (instantaneous envelope and frequency) and 4 additional features: MIF, entropy, spectrum and mean frequency of LF filtered speech signal were merged to the feature vector; 3784 features of speech rhythm are returned in total by the new advanced models.

5. Results

The goal of our experiments is to evaluate the relevance of both taxonomical and new advanced models of speech rhythm for the emotion recognition. In order to compare these rhythmic feature sets with the prosodic one, we used the same classification strategy as in the challenge baseline: we used Support Vector Machine (SVM) classification with linear Kernel, Sequential Minimal Optimisation (SMO) for learning and optimised the complexity on the development partition. The SMO implementation in the WEKA toolkit was used. Because the rhythmic features are numerous and include some redundancy, especially for the new advanced models, we used a correlation-based feature selection method (CFS). This technique reduces the feature space by keeping the features that are highly correlated with the emotional classes while having low inter correlation.

When CFS is performed on all available features of speech rhythm, only 11 of them are kept for Activation and Valence, 8 for Expectation and 6 for Power. In the mean, 45% of these features are derived from the *pPVI* and *PHD* metrics, 33% from the LF signal, 12% from the HHT and the last 10% from the taxonomical models. This result shows the relevance of using dynamic to characterize emotion from speech on two different sides: local dynamic (*pPVI* and *PHD*) and global dynamic (LF). Indeed, local information on both temporal and prosodic shape changes over consecutive words are captured by the *pPVI* and *PHD* metrics, whereas the spectrum of LF signal provides global information on the envelop of speech rhythm. Results also show that using local information (i.e., interval duration between words) to estimate the envelop and the frequency of speech rhythm (HHT) appears to be much less relevant than using global information (i.e., spectrum of LF signal). Results obtained by the rhythmic features in the emotion recognition task, as well as those from the baseline (2 k brute-forced acoustic features [21]), are given in Table 3 for each affective dimension.

In the results, the speech rhythm features outperform the baseline for Valence. The relative improvement obtained by the taxonomical features is equal to 27% for the develop partition and 16% for the test partition (UA measure); 28% and 4% respectively for the advanced features. On the other hand, results are much more contrasted for the three others dimensions. The rhythmic features achieve a lower performance than the baseline for Activitation and both develop and test partitions, excepted with the WA measure and the advanced models for test. The best scores are achieved by the rhythmic features for Expectation on the develop partition, whereas the baseline performs best on the test partition. Concerning Power, the speech rhythm features provide the best performance in all cases excepted for the test partition with the WA measure. In the mean, the baseline achieves best performance for the test partition and the speech rhythm features for develop with the UA measure.

Table 3: Results on the AVEC 2011 Audio Sub-Challenge by the competition measure accuracy for speech rhythm and baseline features. WA stands for weighted accuracy, UA for unweighted accuracy.

Accuracy	ACTIVATION		EXPECTATION		POWER		VALENCE		ALL	
[%]	WA	UA	WA	UA	WA	UA	WA	UA	WA	UA
Taxonomical features										
Develop	43.5	52.4	66.7	66.8	67.7	67.2	66.9	66.9	61.2	63.3
Test	54.8	53.1	47.9	50.0	26.2	41.4	50.7	**54.8**	44.9	49.8
Advanced features										
Develop	45.2	51.0	67.3	67.3	66.9	67.1	67.2	67.9	61.7	63.3
Test	57.6	54.1	51.1	53.0	20.0	**50.0**	46.5	49.2	43.8	51.6
Baseline features										
Develop	63.7	64.0	63.2	52.7	65.6	55.8	58.1	52.9	62.7	56.4
Test	55.0	**57.0**	52.9	**54.5**	28.0	49.1	44.3	47.2	45.1	**52.0**

As a conclusion, results obtained by the features of speech rhythm outperform the prosodic features for emotion recognition for Valence, and achieve much more contrasted results on the others dimension. While there exists difference in the classification results between the two types of features, more detailed future investigation needs to be carried out to understand the relationship between these two types of measures and each emotional dimension. Also note that even if the taxonomical models of speech rhythm have been proposed with a rather different goal in mind, i.e., to quantify cross-linguistic difference rather than intra-language variation due to emotion, they still achieve good performance with best scores in some specific cases.

6. Conclusion

Both taxonomical and novel advanced models of speech rhythm were used for the spontaneous emotion recognition on the recent Audio/Visual Emotion Challenge task, which includes 3.6 hours of natural human affective conversational speech. The performance of these feature sets was compared with the usual prosodic features set. The rhythmic features achieve a promising relative improvement of 16% for Valence, whereas the performance is more contrasted for the three others dimensions. This study thus shows for the first time the relevance of using advanced models of speech rhythm for the characterisation of emotional correlates from speech, especially for Valence. Future work will use specific acoustic anchors of speech (e.g., automatically detected pseudo-phonemes [23]) to provide different structural bases for the metrics of speech rhythm, as well as fusion techniques to estimate their complementarity in the emotion recognition task.

7. References

[1] F. Cummins, "Speech rhythm and rhythmic taxonomy," in *Speech Prosody*, Aix-en-Provence, France, 2002, pp. 121–126.

[2] S. Tilsen, "Multitimescale dynamical interactions between speech rhythm and gesture," *Cogn. Sci.*, vol. 33, pp. 839–879, 2009.

[3] J. R. Evans and M. Clynes, *Rhythm in psychological, linguistic, and musical processes*. Springfield, Charles C. Thomas, 1986.

[4] P. Fraisse, *Les structures rythmiques : étude psychologique*. Publications Universitaires de Louvain, 1956.

[5] E. Zwicker and H. Fastl, *Psychoacoustics: facts and models*. Springer-Verlag, Heidelberg, 1990.

[6] D. Gibbon and U. Gut, "Measuring speech rhythm," in *Eurospeech*, Aalborg, Denmark, 2001, pp. 95–98.

[7] K. R. Scherer, "Psychological models of emotion," *The neuropsychology of emotion*, pp. 137–162, 2000.

[8] J. Ang, R. Dhillon, E. Schriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *Interspeech, 7th ICSLP*, Denver (CO), USA, 2002, pp. 67–79.

[9] V. Dellwo, "The role of speech rate in perceiving speech rhythm," in *Speech Prosody*, Campinas, Brasil, 2008, pp. 375–378.

[10] F. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," *Cognition - International Journal of Cognitive Science*, pp. 265–292, 1999.

[11] M. C. Brady and R. F. Port, "Speech rhythm and rhythmic taxonomy," in *16th ICPhS*, Saarbrcken, Germany, 2006, pp. 337–342.

[12] V. Dellwo, "Rhythm and speech rate: A variation coefficient for ΔC," in *Lang. and Lang. Proc., 38th Ling. Colloq.*, Piliscsaba, Hungary, 2006, pp. 231–241.

[13] E. Grabe and E. Low, "Durational variability in speech and the rhythm class hypothesis," *Papers in laboratory phonology VII*, vol. 7, pp. 515–546, 1977.

[14] L. M. Smith, "A multiresolution time-frequency analysis and interpretation of musical rhythm," Ph.D. dissertation, University of Western Australia, 2000.

[15] F. Lerdahl and R. Jackendoff, *A generative theory of tonal music*. MIT Press, 1996.

[16] S. Tilsen and K. Johnson, "Low-frequency Fourier analysis of speech rhythm," *J. of Acoust. Soc. of Amer.*, vol. 124, no. 2, pp. 34–39, 2008.

[17] F. Ringeval and M. Chetouani, "Hilbert-Huang transform for non-linear characterization of speech rhythm," in *NOLISP*, Vic, Spain, 2009.

[18] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. of Acous. Soc. of America*, vol. 95, pp. 1053–1064, 1994.

[19] H. Xie and Z. Wang, "Mean frequency derived via Hilbert-Huang transform with application to fatigue EMG signal analysis," *Computer Methods and Programs in Biomedicine*, vol. 82, no. 2, pp. 114–120, 2006.

[20] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, L. Noam, A. Kessous, and V. Aharonson, "The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals," in *Interspeech*, Antwerp, Belgium, 2007, pp. 2253–2256.

[21] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "AVEC 2011 — The first international audio/visual emotion challenge," *D'Mello, S. et al. (eds.), ACII 2011, LNCS 6975*, pp. 415–424, 2011.

[22] F. Eyben, M. Wöllmer, and B. Schuller, "The Munich versatile and fast open-source audio feature extractor," in *ACM Multimedia (MM)*, Florence, Italy, 2010, pp. 1459–1462.

[23] F. Ringeval and M. Chetouani, "A vowel based approach for acted emotion recognition," in *Interspeech*, Brisbane, Australia, 2008, pp. 2763–2766.