Rhythm-Based Syllabic Stress Learning without Labelled Data

Bogdan Ludusan[1], Antonio Origlia[2], and Emmanuel Dupoux[1]

[1] LSCP, EHESS/ENS/CNRS, Paris, France
[2] PRISCA-Lab, Federico II University, Naples, Italy

Abstract. We propose a method for syllabic stress annotation which does not require manual labels for the learning process, but uses stress labels automatically generated from a multiscale model of rhythm perception. The model outputs a sequence of events, corresponding to the sequences of strong-weak syllables present in speech, based on which a stressed/unstressed decision is taken. We tested our approach on two languages, Catalan and Spanish, and we found that a classifier employing the automatic labels for learning improves performance over the baseline for both languages. We also compared the results of this system with those of an identical learning algorithm, but which employs manual labels for stress, as well as to the results of a clustering algorithm using the same features. It showed that the system employing automatic labels has a performance close to the one using manual labels, with both classifiers outperforming the clustering algorithm.

Keywords: prosody learning, rhythm, stress annotation

1 Introduction

With the development of speech-based applications most of which use supervised learning principles, there is a growing need for annotations in large quantities. However, manual speech annotation is a long and demanding process. In order to overcome this problem, researchers can employ automatic labelling methods, but even those generally need labelled data to build their model. This becomes more important when the annotation needed is a prosodic one, as there are only a few large corpora annotated for prosody.

The majority of systems proposed for prominence detection use supervised learning and, thus, require manual labels (e.g. [9], [3], [22]), but there are also approaches that do not demand any labelled data, or only partially labelled data. They include methods employing semi-supervised learning (e.g. [15], [12]), using unsupervised learning (e.g. [2], [4]), or rule-based systems (e.g. [24], [1]). Still, many of the above-mentioned studies that employ learning paradigms rely on lexical or syntactic information and, thus, require additional annotations. For this reason, it is desirable to use systems based on acoustic features, as these features either do not require any annotations or, if segmental information is needed for their extraction, it can be obtained relatively cheaply. Thus, we envisage here an

annotation system that exploits the following characteristics: a learning-based approach requiring no labelled examples and using only acoustic features.

Since speech rhythm can be seen as an alternating pattern of strong-weak events, one can employ rhythm information towards learning the stressed/unstressed discrimination. As one of the main prosody components, rhythm has seen an increasing applicative interest recently. Rhythm information has been used for language discrimination/identification [14, 26], syllable nuclei detection [30], speech rate estimation [10], prominence detection [16] or even emotion recognition [21].

We propose here a stress annotation method which takes advantage of the information given by a multiscale model of rhythm perception [28]. The output of the rhythm model has been used already for prominence detection, in conjunction with pitch information [16], and in this paper we employ labels derived from the aforementioned model, to train classifiers. We will show that automatically derived rhythm-based labels are useful for the stress annotation task and that they give a similar performance to when manual stress labels are employed, eliminating the need for prior prosodic annotation.

The paper is structured as follows: in Section 2 we introduce the rhythm perception model employed in this study and we explain how stress labels are obtained from this model. After presenting the materials used and the acoustic features extracted from them, we illustrate, in Section 3, the results obtained with the proposed approach. Section 4 includes an extended discussion of the results and draws some conclusions about the implications of this work.

2 Methods

For the conducted experiments, we used syllable stress labels coming from a model of rhythm perception and we performed classification based on them. This was compared with the results obtained with the same classifier, but using manual stress labels and with a clustering method, on two different languages: Catalan and Spanish. The acoustic features extracted have been previously employed for the task of prominence detection [5], being the best set of features obtained, among the ones considered in that study.

2.1 Rhythm-Based Automatic Label Generation

The stress labels were automatically generated from a model of rhythm perception [28]. The original model takes into account the effect of the peripheral auditory system on the speech signal, by approximating the transformations that the signal undergoes in the human ear. At the output of this pre-processing step the sum of the response of the auditory nerves is obtained, which is further used in the model. The simulation of the auditory nerve response is then processed with a bank of time-domain Gaussian filters having a range of filter widths. For each filter output, the peaks of the function are determined and plotted in a two-dimensional space (time where they occur versus the filter width). By plotting

the peaks for all the filter outputs in this 2D space we would obtain a hierarchical representation, called the rhythmogram, which represents the stronger acoustic events present in speech by peaks over more filter widths.

We employ in this paper an approximation of the original model, which was previously used for prominence detection [16]. The approximation uses the signal energy rather than the output of the nerve response, the former being more convenient for computing the rhythmogram, in the case of large scale events [29]. Furthermore, the output of the model was quantized, the rhythmogram being represented by a sequence of events and their heights (called P-values). The P-value of each event is obtained by summing all peak values of the event, across all filter widths (see [14] for more details).

The parameters required by the models (minimum and maximum filter width, and the total number of filters in the filter bank) were the ones obtained in [16], on a small corpus of English speech. The following steps were adopted for obtaining the rhythmogram: the speech signal was resampled at 500 Hz and full wave rectification was performed on it. The ear's loudness function was then modelled, by taking the cubic root of the signal, and one hundred logarithmically-distanced Gaussian filters were applied. The resulting representation is illustrated in Figure 1 for one of the sentences used in the experiments. The upper part shows the waveform of the speech signal, the middle part the obtained rhythmogram, while the lower part displays the syllable-level segmentation along with the oracle syllabic stress annotation.

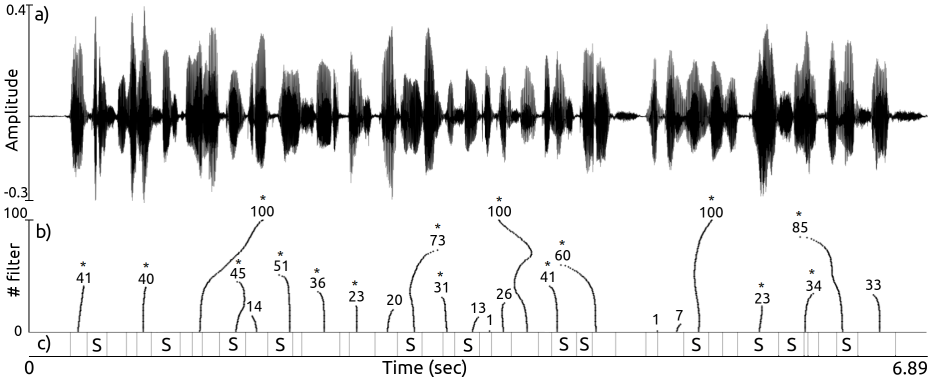

Fig. 1. *a) Waveform of the phrase "El precio de la cesta de la compra puede ser hasta un veinticinco por ciento más caro, dependiendo de cuál sea el comercio en el que compremos"; b) rhythmogram of the phrase, along with the corresponding P-values (values marked with an asterisk represent local maxima of the stress function); and c) oracle syllable segmentation and stress annotation (label S represents stressed syllables, while no label corresponds to unstresssed syllables).*

Taking a closer look at the rhythmogram in Figure 1 we can see that it contains 24 events (the lines in the middle panel), with P-values ranging from

1 to 100. The points plotted at the bottom level represent the peaks obtained with the narrowest filter from the bank of filter (filter index 1), thus the output function contained many peaks (24). As the filter index increases, the applied filter becomes wider and smooths more the signal energy, thus, when reaching index 60 the output of the filter would have only 6 peaks. The maximum index, 100, corresponds to the widest filter applied and, for this particular example, it gave 3 peaks in its output.

In order to obtain syllable-level stress labels, we first compute the P-values for all the events of the rhythmogram. Then, we determine the events between the boundaries of each syllable and take their P-value. If multiple events are found, only the largest event (in P-value) is considered. If no event was found, the value 0 is assigned to that particular syllable. Next, we define a stress function with the obtained P-values for all the syllables and we consider the local maxima of this function as being the stressed syllables, with the rest of the syllables belonging to the unstressed class. The local maximum is computed in a three-syllable window, centered on the current syllable. For this study we used the syllable boundaries available with the corpus, the syllable segmentation being obtained from automatically aligned phonetic transcription.

Going back to the example illustrated in Figure 1, the resulting stress function will have its first value equal to 40, as the P-value of the event found within the boundaries of the first syllable was 40. The following three values will be equal to 0, since no rhythmogram event was found inside the following three syllables, followed by 38 (the height of the second event, which falls inside syllable number five). Next, we have two other values equal to 0, followed by 100 (the height of the third rhythmogram event, which falls inside the eight syllable of the utterance), and so on. Please note that in order to assign an event as belonging to a particular syllable it has to 'begin' inside that syllable (the peak obtained with the narrowest width should fall inside the syllable boundaries).

2.2 Acoustic Features

The acoustic features set used in the experiments is the one that obtained the best performance in [5]. The considered features are:

- syllable length (*sylLen*)
- nucleus length, estimated as the -3dB band of the energy maximum inside the syllable (*nucLen*)
- average energy inside the syllable nucleus (*avgEne*)
- ratio between the voiced time in the syllable and the total syllable length (V_L)
- likelihood of hearing the pitch movement passing through the nucleus as a dynamic tone (glissando) (Γ)

This last parameter is motivated by previous studies on tonal movement perception [23, 27, 11, 18] and has also been used in studies on pitch stylization [19, 20]. After producing a linear stylization of the pitch profile with the approach

presented in [20] and considering the segment $[s_1, s_2]$ passing through the syllable nucleus, the Γ_{s_1,s_2} parameter is then computed as follows:

$$\Gamma_{s_1,s_2} = \begin{cases} 1 & if\ V_{s_1,s_2} > \frac{0.32}{T_e^2} \\ \\ \frac{V_{s_1,s_2} T_e^2}{0.32} & otherwise \end{cases} \quad (1)$$

where V_{s_1,s_2} is the absolute rate of change of the pitch excursion (in semitones/second). The glissando perception threshold $\frac{0.32}{T_e^2}$ used as reference in this formula is the same as the one used in [18]. While a tonal movement exceeding the threshold will have value 1, movements below it will be given a glissando likelihood value between 0 and 1. The Γ_{s_1,s_2} parameter reacts both to rising and descending pitch movements.

2.3 Learning Algorithms

For stress classification we have chosen a simple learning algorithm, naive Bayes. It is a probabilistic linear classifier which assumes that feature values are independent of other features, given the class. Since part of the features employed are not normally distributed (especially V_L and Γ), we decided to employ kernel-density estimate using Gaussian kernels [13]. The implementation of the algorithm is the one given by the Weka toolbox [8].

In order to compare the performance of the classifier that uses automatic labels with that of a system employing no annotated labels, we have chosen to use an unsupervised method based on the expectation-maximization (EM) algorithm. The EM clustering models the features by means of a mixture of Gaussians, the particular implementation used here [8] considering a similar feature independence assumption to the naive Bayes classifier.

2.4 Materials

We use for the experiments presented here the news sub-part of the Glissando corpus [7]. The chosen part contains broadcast news recorded by professional speakers in two languages: Catalan and Spanish, totalling over 12 hours of speech (equally divided between the two languages). For each language the same number of speaker, 8, was employed, 4 females and 4 males.

The corpus was entirely transcribed and annotated, including segmental and prosodic levels. At the segmental level, both phonetic and syllabic segmentations are available, while the prosodic level contains prosodic boundary annotations. Besides prosodic phrasing, the corpus is also marked for syllabic stress, which we will use in our experiments as oracle labels.

3 Experiments

The experiments were conducted in a leave-one-speaker-out cross-validation setting. By training the model on all the speakers except one and testing on

the latter, we ensure a speaker-independent model that does not capture any individual speech particularities.

For the evaluation of the results we employed the area under the receiver operating characteristic (ROC) curve (AUC). The ROC curve is obtained by varying the class decision threshold and plotting the obtained true positive rate (TPR) versus false positive rate (FPR), for each value of the threshold. We use the AUC due to its straightforward way in interpreting the results, as well as for being a measure independent of the size and the distribution of the label classes. Given a random pair of syllables, one representing a stressed syllable, and the other an unstressed syllable, the AUC can be described as the probability of making a correct choice in a forced choice task and thus the chance level for the AUC is always equal to 0.5. Furthermore, as we are interested in the overall learning quality, the AUC provides a single value for both classes (stressed/unstressed), whereas other measures, like precision, recall or F-score, would need two values to characterize the system, one for each class. While both oracle and automatically generated labels were used in the learning process, the evaluation was always performed using the oracle stress labels as the reference.

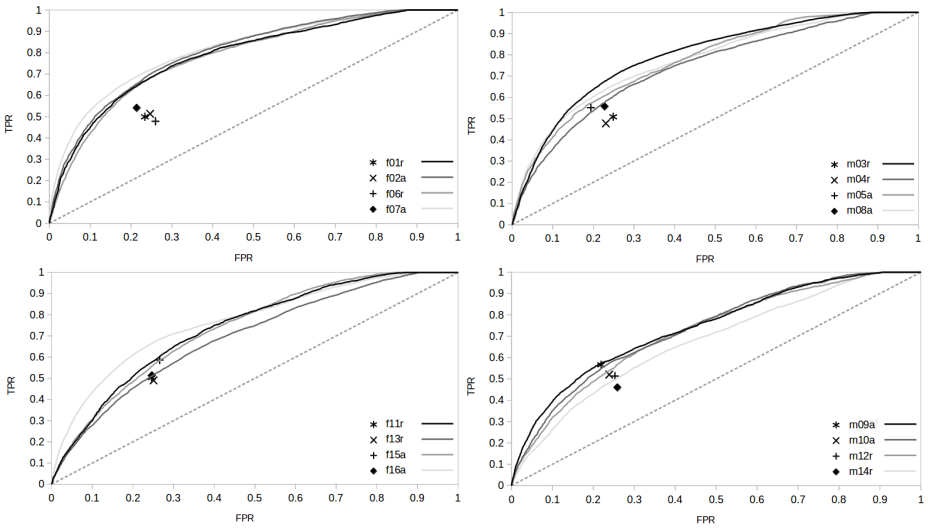

Fig. 2. *Classification results obtained when labels derived from the rhythmogram are used for learning a naive Bayes classifier on acoustic features (curves in the ROC space), compared with the results given by the baseline (points in the ROC space), for each speaker. The upper panel shows the results for Catalan, while the lower panel the ones for Spanish. For the two languages, the results corresponding to the female speakers are on the left and those for the male speakers on the right. Chance level is represented by a dotted line.*

3.1 Learning with Automatic Labels

The rhythm-based procedure presented in Section 2.1 was applied in order to obtain the automatic labels used in the learning process and it was also considered as the baseline (further called *baselineRhy* system). The results obtained with the *baselineRhy* and the supervised algorithm that employs automatic labels for learning (referred to as the *rhyLabel* system) are displayed in Figure 2. They are illustrated for each speaker of the two languages investigated (Catalan in the upper panel and Spanish in the lower panel), with female speakers on the left and male speakers on the right side. Each speaker is represented by a point in the FPR-TPR space, in the case of the *baselineRhy* and by a curve in the same space, for the *rhyLabel* system.

They show that stress learning based on labels derived from a rhythm model improves over the baseline (the model used to generate the labels, in the first place). It appears though that there are some differences between languages: while for Catalan we obtained important improvements for all speakers, we had lower improvements for Spanish, with one speaker seeing no improvement over the baseline. If we are to compare the *baselineRhy* results with those of the *rhyLabel* system, by taking, for each speaker, the point on the curve which corresponds to the same FPR as the baseline, we would obtain: an increase in TPR, for Catalan, ranging between 2.2% and 21.8% (14.0%, on average), and a change in TPR, for Spanish, ranging between -0.1% and 15.1% (5.0%, on average).

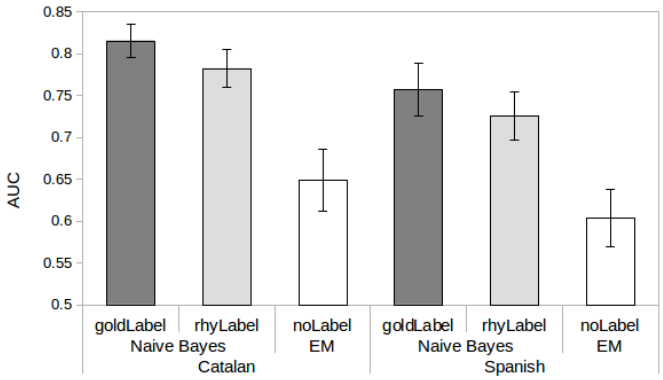

Fig. 3. *Area Under the ROC Curve classification results obtained for Catalan and Spanish, when using the following systems: naive Bayes employing oracle labels (*goldLabel*), naive Bayes employing labels obtained from the rhythmogram (*rhyLabel*) and EM clustering (*noLabel*). The error bars represent the standard deviation across speakers. The three systems use the same set of acoustic features.*

Next, we compared the results obtained when automatic labels are used to learn a classifier (*rhyLabel*) to those obtained when manual stress annotations are employed with the same learning algorithm and the same acoustic features

(*goldLabel*). These two cases were then compared against a clustering algorithm that takes in input the same features (*noLabel*). Both comparisons (*rhyLabel* $\leftrightarrow$ *goldLabel* and *rhyLabel* $\leftrightarrow$ *noLabel*) are important, as the former would show how close the use of automatic labels is to the use of manual labels, while the latter would indicate whether the proposed approach performs better than another system that does not employ manually labelled data for learning. Figure 3 displays the comparison, for both languages. Per-speaker two-tailed paired t-tests were used to test the significance of the difference between the *rhyLabel* $\leftrightarrow$ *goldLabel* and *rhyLabel* $\leftrightarrow$ *noLabel* and they showed that all differences are statistically significant, for both languages ($p < 0.001$).

As expected, the best performance, on both languages, is obtained by the classification system employing manual labels. The classifier making use of automatically determined labels performs well, with the difference between it and the topline being only about 3% absolute value, for both Catalan and Spanish. Furthermore, its performance is well above that of the clustering method (13.3% and 12.2% absolute difference for Catalan and Spanish, respectively). This result demonstrates that learning without labels can be improved, being able to reach performances relatively close to that of a system employing manual annotations.

3.2 Learning with Different Classifiers

In order to test the generalizability of the results obtained, we trained two additional classifiers, one based on logistic regression and the other one based on support vector machines (SVMs), using the same setting as in the previous experiment (same acoustic features, trained with both manual and automatic labels). The logistic regression classifier had no parameters to set, so the training and evaluation procedures were identical those of the naive Bayes classifier. For SVMs instead, we used a Radial Basis Function kernel, thus the tuning of the C and γ parameters was needed. To optimize them, for each speaker independent test session, a subset of the training set (10%) was automatically extracted using stratified sampling to keep the class distribution consistent. This subset was used to optimize the parameters by training an SVM on 70% of the selected material and testing on the remaining 30%. The search space was limited to the interval $[0.5 - 1.5]$ and an exhaustive grid (step size = 0.1) was used to select the values of the considered parameters yielding the highest AUC. These parameters were then used to train a model on the full training set. The evaluation was performed on the left-out speaker which never appeared, neither in the training phase nor in the optimization phase.

When we compare the results obtained with the classifiers trained with the rhythmogram-derived labels with those of the rhythmogram system alone we observe significant improvements (see Table 1). The TPR increase when using logistic regression and SVMs, for an equivalent FPR, reaches 8.4% and 10.2% for Catalan and 0.6% and 4.7% for Spanish. The *rhyLabel* $\leftrightarrow$ *goldLabel* comparison (in Table 2) shows an advantage of 5.3% and 6.4% (Catalan) or 4.7% and 2.9% (Spanish), for the latter. The *rhyLabel* $\leftrightarrow$ *noLabel* comparison shows the same trend as the naive Bayes classifier: an improvement over the unsupervised learner

Table 1. Average improvement in TPR, for an equivalent FPR level, between the baselineRhy and rhyLabel systems, for the three classifiers used in this study.

Learning algorithm	Catalan	Spanish
Naive Bayes	.140	.050
Logistic Regression	.084	.006
Support Vector Machines	.102	.047

of 11.7% and 8.5%, for Catalan, or 10.7% and 8.6%, for Spanish. SVMs appear to perform slightly worse than naive Bayes and logistic regression: a possible explanation for this is that SVMs are more sensitive to low-quality annotations than the other considered approaches [6]. The performance difference between the learning algorithms and the two languages will need to be investigated further.

Table 2. AUC results obtained with the rhyLabel and goldLabel system when employing different learning algorithms. For comparison, we report also the results of the noLabel system, which uses EM clustering.

Learning algorithm	Catalan			Spanish		
	goldLabel	rhyLabel	noLabel	goldLabel	rhyLabel	noLabel
Naive Bayes	.815	.782	-	.757	.726	-
Logistic Regression	.819	.766	-	.758	.711	-
Support Vector Machines	.798	.734	-	.719	.690	-
EM clustering	-	-	.649	-	-	.604

4 Discussion and Conclusions

We presented here an approach for stress annotation that does not employ any manual labels for the learning process. We have shown that the labels obtained from a rhythm-based system, based only on the speech signal, can be successfully employed for learning. The trained classifier outperformed both the baseline and a clustering system that used the same acoustic features, while giving a performance close to that of a classifier that employs manual prosodic labels and identical features. As seen also in [16], the information given by the rhythmogram seems to be quite robust and language independent. In the aforementioned paper, the parameters of the system were determined on English and then tested on Italian and French, with good results. We applied here the same parameter settings on two new languages, Spanish and Catalan, with equally good performance.

The current study can be seen as one step further in improving the automatic annotation of prosody, without the use of labelled data in the learning process. With more improvements in this direction we can envisage more prosodic studies and better analyses, as the access to annotated data increases. This could be especially important as a starting point for languages which, for the moment, have no prosodic annotations. At the same time, an increase in available prosodic annotation might lead to more speech applications taking advantage of prosodic knowledge or to improvements in systems which use prosody determined through unsupervised or rule-based methods [17].

As mentioned before, the approach proposed here employs no labelled data in the learning process. Still, it made use of a small prominence annotated dataset for the setting of the parameters of the rhythmogram. But since those parameters were determined on English and our approach tested on two different languages, Catalan and Spanish, we could assume that the same procedure can be applied to other languages also. By making use of less than 20 minutes of annotation from a highly resourced language like English, we were able to annotate with reasonable performance more than 12 hours of data from two previously unseen languages.

One significant issue that remains to be investigated is the role that syllable segmentation plays on the stress annotation performance. Since none of the current signal-based methods for syllable segmentation give excellent results, it is important to know to what extent the segmentation errors affect the prosodic annotation learning. One alternative would be to derive syllables from forced-aligned phonemic annotations. While this would greatly reduce the number of languages on which the approach could be applied, there are many more languages that have enough data to train acoustic models for forced alignment than languages in which extensive prosodic annotations exist. If this is not an option and a manual annotation has to be performed, it is nevertheless cheaper and easier to perform the segmentation stage manually (since this would also be needed to further add the prosodic annotation) and to have a system that would do a first pass prosodic annotation that needs to be corrected, than to do the whole prosodic annotation from scratch.

Our investigation was limited to the annotation of stressed syllables because of the lack of large corpora annotated for prominence. Once such corpora become available we plan to extend our study to the annotation of the more general phenomenon of acoustic prominence. Since the system employed in this study for the automatic generation of stress labels has been used successfully for prominence detection, we expect the procedure proposed here to perform well also in the case of acoustic prominence annotation.

We considered the present investigation as a case study on the use of automatically generated labels for learning stress and, thus, the emphasis was not put on the learning algorithms. In order to obtain better performance we would like to explore the use of state-of-the-art learning paradigm for prominence labelling, latent dynamic conditional neural fields [25], as well as learning algorithms which are robust to low quality labels [6].

Acknowledgments. BL and ED's work was funded by the European Research Council (ERC-2011-AdG-295810 BOOTPHON). It was also supported by the Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*), the Fondation de France, the École des Neurosciences de Paris, and the Région Île-de-France (DIM cerveau et pensée). AO's work was supported by the Italian Ministry of University and Research and EU under the PON OR.C.HE.S.T.R.A. project.

References

1. Abete, G., Cutugno, F., Ludusan, B., Origlia, A.: Pitch Behavior Detection for Automatic Prominence Recognition. In: Proceedings of Speech Prosody (2010)
2. Ananthakrishnan, S., Narayanan, S.: Combining Acoustic, Lexical, and Syntactic Evidence for Automatic Unsupervised Prosody Labeling. In: Proceedings of INTERSPEECH. pp. 297–300 (2006)
3. Ananthakrishnan, S., Narayanan, S.: Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence. Audio, Speech, and Language Processing, IEEE Transactions on 16(1), 216–228 (2008)
4. Chiang, C.Y., Chen, S.H., Yu, H.M., Wang, Y.R.: Unsupervised Joint Prosody Labeling and Modeling for Mandarin Speech. The Journal of the Acoustical Society of America 125(2), 1164–1183 (2009)
5. Cutugno, F., Leone, E., Ludusan, B., Origlia, A.: Investigating Syllabic Prominence With Conditional Random Fields and Latent-Dynamic Conditional Random Fields. In: Proceedings of INTERSPEECH. pp. 2402–2405 (2012)
6. Folleco, A., Khoshgoftaar, T., Van Hulse, J., Bullard, L.: Identifying Learners Robust to Low Quality Data. In: Proceedings of the IEEE International Conference on Information Reuse and Integration. pp. 190–195 (2008)
7. Garrido, J.M., Escudero, D., Aguilar, L., Cardeñoso, V., Rodero, E., De-La-Mota, C., González, C., Vivaracho, C., Rustullet, S., Larrea, O., et al.: Glissando: A Corpus for Multidisciplinary Prosodic Studies in Spanish and Catalan. Language resources and evaluation 47(4), 945–971 (2013)
8. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. SIGKDD Explorations 11(1), 10–18 (2009)
9. Hasegawa-Johnson, M., Chen, K., Cole, J., Borys, S., Kim, S.S., Cohen, A., Zhang, T., Choi, J.Y., Kim, H., Yoon, T., et al.: Simultaneous Recognition of Words and Prosody in the Boston University Radio Speech Corpus. Speech Communication 46(3), 418–439 (2005)
10. Heinrich, C., Schiel, F.: Estimating Speaking Rate by Means of Rhythmicity Parameters. In: Proceedings of INTERSPEECH. pp. 1873–1876 (2011)
11. House, D.: Differential Perception of Tonal Contours Through the Syllable. In: Proceedings of ICSLP. pp. 2048–2051 (1996)
12. Jeon, J.H., Liu, Y.: Semi-Supervised Learning for Automatic Prosodic Event Detection Using Co-Training Algorithm. In: Proceedings of ACL-IJCNLP. pp. 540–548 (2009)
13. John, G., Langley, P.: Estimating Continuous Distributions in Bayesian Classifiers. In: Proceedings of the Eleventh conference on Uncertainty in artificial intelligence. pp. 338–345 (1995)

14. Lee, C., Todd, N.M.: Towards an Auditory Account of Speech Rhythm: Application of a Model of the Auditory "Primal Sketch" to Two Multi-Language Corpora. Cognition 93(3), 225–254 (2004)
15. Levow, G.A.: Unsupervised and Semi-Supervised Learning of Tone and Pitch Accent. In: Proceedings of NAACL-HLT. pp. 224–231 (2006)
16. Ludusan, B., Origlia, A., Cutugno, F.: On the Use of the Rhythmogram for Automatic Syllabic Prominence Detection. In: Proceedings of INTERSPEECH. pp. 2413–2416 (2011)
17. Ludusan, B., Ziegler, S., Gravier, G.: Integrating Stress Information in Large Vocabulary Continuous Speech Recognition. In: Proceedings of INTERSPEECH. pp. 2642–2645 (2012)
18. Mertens, P.: The Prosogram: Semi-Automatic Transcription of Prosody Based on a Tonal Perception Model. In: Proceedings of Speech Prosody (2004)
19. Origlia, A., Abete, G., Cutugno, F.: A Dynamic Tonal Perception Model for Optimal Pitch Stylization. Computer Speech and Language pp. 190–208 (2013)
20. Origlia, A., Cutugno, F.: A Simplified Version of the OpS Algorithm for Pitch Stylization. In: Proceedings of Speech Prosody. pp. 992–996 (2014)
21. Ringeval, F., Chetouani, M., Schuller, B.W.: Novel Metrics of Speech Rhythm for the Assessment of Emotion. In: Proceedings of INTERSPEECH. pp. 346–349 (2012)
22. Rosenberg, A.: AutoBI - A Tool for Automatic ToBI Annotation. In: Proceedings of INTERSPEECH. pp. 146–149 (2010)
23. Rossi, M.: Interactions of Intensity Glides and Frequency Glissandos. Language and speech 21(4), 384–396 (1978)
24. Tamburini, F.: Automatic Prominence Identification and Prosodic Typology. In: Proceedings of INTERSPEECH. pp. 1813–1816 (2005)
25. Tamburini, F., Bertini, C., Bertinetto, P.M.: Prosodic Prominence Detection in Italian Continuous Speech Using Probabilistic Graphical Models. In: Proceedings of Speech Prosody. pp. 285–289 (2014)
26. Tepperman, J., Nava, E.: Long-Distance Rhythmic Dependencies and Their Application to Automatic Language Identification. In: Proceedings of INTERSPEECH. pp. 1061–1064 (2011)
27. t'Hart, J., Collier, R., Cohen, A.: A Perceptual Study of Intonation: An Experimental-Phonetic Approach. Cambridge University Press, Cambridge (1990)
28. Todd, N.M.: The Auditory "Primal Sketch": A Multiscale Model of Rhythmic Grouping. Journal of New Music Research 23(1), 25–70 (1994)
29. Todd, N.M., Brown, G.: Visualization of Rhythm, Time and Metre. Artificial Intelligence Review 10, 253–273 (1996)
30. Zhang, Y., Glass, J.R.: Speech Rhythm Guided Syllable Nuclei Detection. In: Proceedings of ICASSP. pp. 3797–3800 (2009)