An Adaptive Neuro-Fuzzy Inference System for the Qualitative Study of Perceptual Prominence in Linguistics

Autilia Vitiello*, Giovanni Acampora†, Francesco Cutugno‡, Petra Wagner§ and Antonio Origlia¶
*Department of Computer Science, University of Salerno, Italy, Email: avitiello@unisa.it
†Department of Physics "Ettore Pancini", University of Naples Federico II, Italy, Email: giovanni.acampora@unina.it
‡Department of Electrical Engineering and Information Technology, University of Naples Federico II, Italy,
Email: cutugno@unina.it
§Faculty of Linguistics and Literature, University of Bielefeld, Germany, Email: petra.wagner@uni-bielefeld.de
¶Department of Information Engineering, University of Padova, Italy, Email: aoriglia@unipd.it

Abstract—**This paper explores the applications of fuzzy logic inference systems as an instrument to perform linguistic analysis in the domain of prosodic prominence. Understanding how acoustic features interact to make a linguistic unit be perceived as more *relevant* than the surrounding ones is generally needed to study the cognitive processes needed for speech understanding. It also has technological applications in the field of speech recognition and synthesis. We present a first experiment to show how fuzzy inference systems, being characterised by their capability to provide detailed insight about the models obtained through supervised learning can help investigate the complex relationships among acoustic features linked to prominence perception.**

I. INTRODUCTION

Similarly to music signals, a physical chain of speech sounds undergoes to rhythmical constraints. However, while the musical sequence takes into account these rhythmical factors just to render accent and timing factors, in speech a further complication is given by the concept that each speech portion is also corresponding to some linguistic unit. This means that when speech portion is rhythmically relevant also the correspondent linguistic unit assumes a special evidence in the sequence. The phenomenon by which a linguistic unit in a sequence, given the acoustic characteristics of the correspondent speech portion, is considered more *important* than others is much debated and the many views it can be analysed from, by a linguistics perspective [1], makes it even difficult to find a common description for it. Prosody is the discipline studying intonational and rhythmical features in speech, and, according to prosody, the temporal units where acoustic properties merge with linguistic structure when rhythmical factors are concerned is the syllable.

As a matter of fact, the most widely accepted definition of prominence is also the most generic one. In the view proposed by [2], a linguistic unit has been defined to be *"[. . .] prosodically prominent when it stands out from its environment (by virtue of its prosodic characteristics)"* [2, p. 89]. Given the difficulties of defining prominence, even convening towards a common annotation protocol has posed serious problems to the scientific community working on the problem. Views considering prominence as a categorical phenomenon are mostly related to a functional view of it, while theoretical approaches treating the subject from a physical point of view have moved towards continuous or quasi-continuous scales. The capability of human raters to manually indicate the degree of prominence of a specific linguistic unit is also questionable. First of all, agreement among judges is typically not high because of the high degree of subjectivity. Also, quasi-continuous scales adopted in the past have been criticised because of the arbitrary definition of the number of levels chosen for the task. To address the last issue, recent work has adopted a gestured based approach to prominence annotation [3].

Given the wide set of perspectives involved in prominence study, it is necessary to specify that our interest, in this work, is to adopt a physical approach to prominence. As such, the weight and interplay of acoustic measures extracted from the recorded speech signal are going to be investigated. It is important to specify, however, that although we concentrate on this specific perspective in this work, interactions with other perspectives are foreseen in future analyses. The concept of prominent units in spoken communication, from a signal-related point of view, has been deeply investigated in the last years but conclusive findings about the way prominence is conveyed have not been provided and language specificity has only briefly been investigated (i.e. [4]). Contribution from speech technology has come, in this field, by proposing algorithms designed to detect, for example, relevant prosodic variations [5], automatically annotate recorded and segmented speech and evaluate the final result by comparing the annotation with a manual reference. Many approaches have investigated the problem using a top-down strategy, trying to evaluate the effect of theoretical assumptions on automatic annotation [6], [7]. Bottom-up approaches using machine learning have, instead, tried to obtain indirect indications about the way prominence is perceived by comparing the performance of similar models exhibiting very specific differences. In [8], the superior performance of Latent-Dynamic Conditional Random Fields with respect to Conditional Random Fields suggested that a latent dynamic was present in the acoustic features of sequences of non-prominent syllables to signal or at least provide a bias towards perceiving the next unit as prominent. Subsequent work [9], showed that Latent Dynamic Conditional Neural Fields systematically outperformed both Latent Dynamic Conditional Random Fields and Conditional Neural Fields, suggesting that the relationship between acoustic measures and observations is non-linear in nature. These kind of approaches, while having the advantage of being data driven, is difficult to interpret as most models obtained with machine learning approaches do not explicitly provide information about the way the different features interact and influence the final result.

In this work, we describe how a data-driven modelling approach using fuzzy inference can provide insight about the way acoustic features interact among each other to create the perception of prominence. Specifically, we train an ANFIS system on a human-annotated dataset of speech utterances and present a linguistic interpretation of the control surfaces provided by the system. We will show that expected interactions are found in the obtained models and we will highlight specific situations where detailed information about features contribution to prominence scoring can be obtained.

II. MATERIAL

The database used for prominence detection is the Bonner Prosodische Datenbank [10]. It consists of sentences and short stories read by 3 native speakers of German. The data has been manually annotated for syllable and boundary prominence by three trained phoneticians based on the procedure described in [11], who described prominence as a continuous rather than a categorical parameter. Prominence was annotated on a continuous scale ranging from 0-31. The inter-labeller agreements were high and their correlations ranged between 0.74 and 0.86. After labelling, the median prominences were calculated out of the three labellers prominence ratings for each syllable. The medians are used as reference values of perceptual prominence in our subsequent experiments. In our experiments, the data related to the first two speakers of the aforementioned database (below Speaker1 and Speaker2) will be used.

III. METHODS

In this section, we present the features extraction procedure and summarise the ANFIS fuzzy system and its application to the considered case study.

A. Features extraction

Syllable structure can vary depending on the phonetic segments involved in speech production. A syllable always has a *nucleus*, an optional consonantic *coda* and an optional consonantic *head*. Heads and codas may contain more than one consonant and syllable grouping is governed by the sonority sequencing principle [12]: segments characterised by higher sonority tend to be closer to the syllable nucleus while less sonorant segments are positioned towards the boundaries. The most sonorant segment, typically the vowel, is the syllable nucleus and, consequently, moving away from the syllable nucleus corresponds to producing segments with lower sonority. As segments sonority rises again, a new syllable starts. This principle is a linguistic universal and is found across languages with few, specific exceptions.

Segments durations, the intensity of the syllable nucleus and pitch behaviour inside the nucleus are the main cues to signal prominence acoustically [13], [14]. While other cues linked, for example, to voice quality are also considered in the literature, in this work we concentrate on this set of acoustic measures to verify that the interpretation of ANFIS control surfaces provides a view that is linguistically interpretable and coherent. Starting from the phone and syllable-level segmentation provided with the considered material, an automatic procedure for acoustic features extraction was designed using the software PRAAT [15].

For each considered utterance, the features extractor cycled through the syllables and selected the phone with the highest intensity as the nucleus, following the sonority sequencing principle. Both the duration of the entire syllable and the duration of the syllable nucleus are considered as features in this work as it is generally not clear whether one is more important than the other or if they interact in some way. Also, from the syllable nucleus the average intensity and the average pitch were extracted. As perceptual prominence is a phenomenon linked to the specific syllable context, a windowing procedure with zero-padding at the extremes was adopted to provide the ANFIS system with this specific knowledge. In this work, we used a one-sized window to present a set of first results, so the acoustic features of the considered syllable and of its immediate neighbours are included in each features vector of the dataset. Previous work [8] has shown that context may extend up to two neighbouring syllables, at least for Italian and English, but this kind of analysis is left for future work.

B. ANFIS

ANFIS is an adaptive network and, as such, it works to achieve a desired input-output by updating parameter sets according to given training data and a gradient-based update procedure [16]. In particular, this updating feature is used in ANFIS to learn and adapt the parameters of a given Takagi-Sugeno-Kang (TSK) fuzzy inference system. As example, Fig. 1 shows an adaptive network that models a first-order TSK fuzzy inference systems composed by two rules:

1) If x is A_1 and y is B_1 Then $f_1 = p_1 \cdot x + q_1 \cdot y + r_1$
2) If x is A_2 and y is B_2 Then $f_2 = p_2 \cdot x + q_2 \cdot y + r_2$

In literature, ANFIS has been successfully used in different and several application domains, such as medical diagnosis [17], economy [18], robotics [19] and image processing [20]. In Speech and Natural Language Processing, ANFIS has been already proposed, for instance, to model the relationship between acoustic features and emotion dimension [21], to predict the imprecise nature of speech prosody [22] and to identify the speaker, language and the words spoken [23]. However, no work exists about the application of ANFIS for a prominence study. Therefore, in this work, we propose ANFIS to perceptual prominence identification for the first time. In this context, the exploitation of ANFIS will allow both to create the perception of prominence and to explicitly provide information about the way the different features interact and influence the final result.

ANFIS works by updating an initial TSK during a set of epochs. In this work, the initial TSK is built through the subtractive clustering algorithm [24]. Starting from this initial

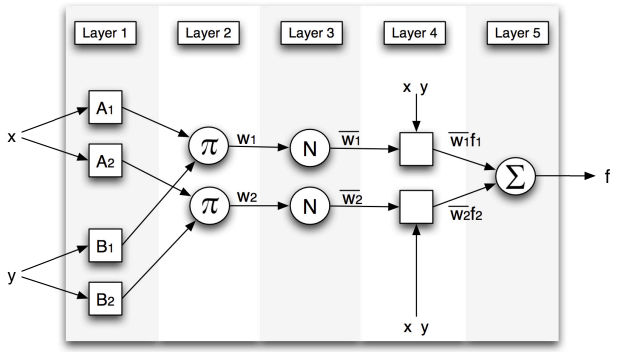

Fig. 1. The Adaptive Neuro Fuzzy Inference System [16]

TSK, ANFIS is executed for 10 epochs by producing a 12-rule TSK fuzzy model. The data used to learn a TSK are represented by the features described in the previous section extracted from the dataset Speaker1. In the next section, the behavior of the trained TSK system is discussed to extract information about how acoustic features interact one with each other.

IV. RESULTS AND DISCUSSION

This section is mainly devoted to perform a qualitative analysis of the TSK system trained through ANFIS in order to extract information about relations between acoustic features and between them and prominence. The section is concluded with a quantitative analysis aimed at showing the goodness of the trained TSK system and supporting the validity of the carried out qualitative analysis. Hereafter, details about this two-fold analysis are given.

A. Qualitative Analysis

The control surface shown in Figure 2 describes an interesting relationship between duration and pitch. While the two often interact, they can lend prominence independently. Duration appears to be able to increase prominence scoring independently of pitch while a minimum value for duration appears to be necessary for higher pitch values to actually contribute to increasing prominence scoring. This makes sense as the human capability of discriminating pitch levels and movements depends on their duration [25], [26]. Figure 3 shows that this effect is stronger when pitch is compared to the nucleus duration, as pitch movements that are relevant for prominence are typically found in the nucleus. High intensity, although yielding higher prominence values when duration is low if compared to pitch, is influenced by duration getting longer, too, as shown in Figure 4.

From the analysis of the control surface presented in Figure 5, duration and intensity seem to work in a more additive fashion: the combination of high values for the two features is needed to reach maximum prominence. The rising surface is also not linear: the way the surface rises suggests that small differences in duration have a different effect in yielding prominence depending on the intensity contribution. For low

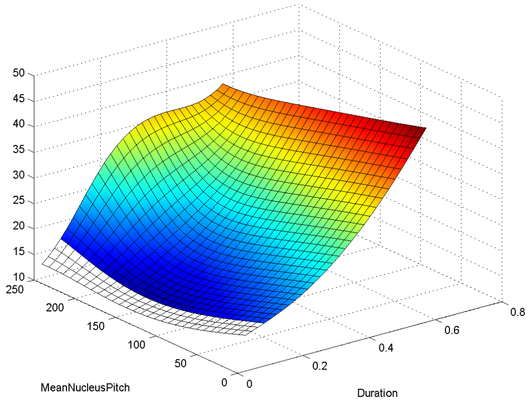

Fig. 2. Control surface for the interaction between Mean Nucleus Pitch and Duration features

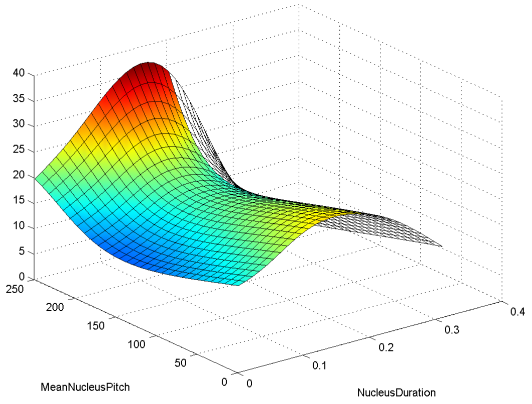

Fig. 3. Control surface for the interaction between Mean Nucleus Pitch and Nucleus Duration features

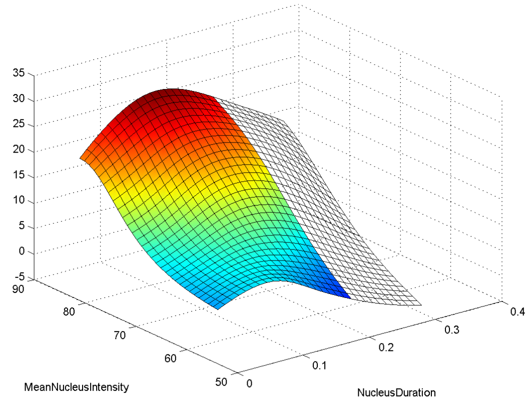

Fig. 4. Control surface for the interaction between Mean Nucleus Intensity and Nucleus Duration features

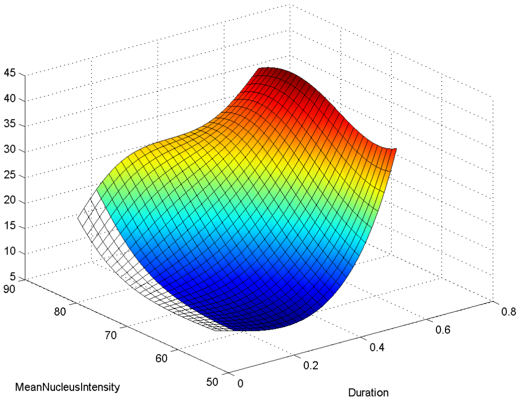

Fig. 5. Control surface for the interaction between Mean Nucleus Intensity and Duration features

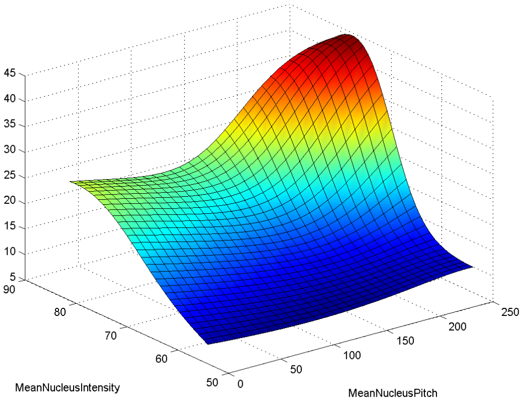

Fig. 6. Control surface for the interaction between Mean Nucleus Intensity and Mean Nucleus Pitch features

intensity, the rise given by duration is exponential while when energy is high, too, the prominence score increases in a more sigmoid-like way. Duration appears to yield a stronger prominence score when intensity is not contributing than what can be observed when duration is short and intensity is high. The control surface in Figure 6 suggests that like duration, intensity can also act independently of pitch to make a syllable prominent. The combination of the two yields higher prominence levels and there appears to be a specific boundary over which prominence is perceived more strongly.

The previous analysis is concerned with the relationships among features belonging to the syllable of interest. As prominence is a phenomenon that is severely influenced by context, an important kind of investigation to conduct concerns the relationship between acoustic features of the syllable of interest and the same features extracted from its neighbouring units. The control surface shown in Figure 7 highlights that longer duration of the preceding syllable has a negative impact on prominence. This is consistent with theoretical expectations as the degree of perceived prominence for the current syllable

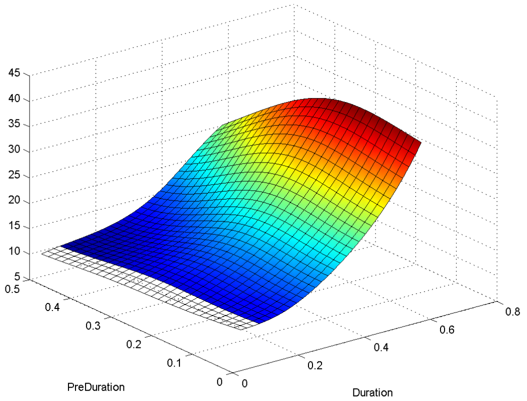

Fig. 7. Control surface for the interaction between the Duration of the preceding syllable and the duration of the current syllable

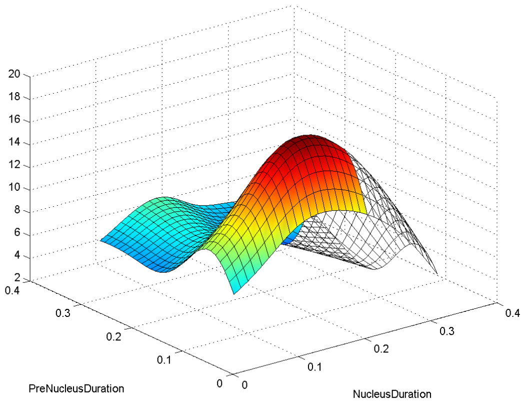

Fig. 8. Control surface for the interaction between the Nucleus Duration of the preceding syllable and the Nucleus Duration of the current syllable

is dampened if the preceding one has longer duration. The same applies for nuclear durations when the comparison is made with the preceding syllable, as shown in Figure 8.

The control surface shown in Figure 9 shows that a syllable is more prominent if it is longer than the following one. In general, duration appears to influence prominence scoring independently of post duration. Consistently with theoretical expectations, a long duration of the syllable of interest matched with a short duration of the following syllable yields a strong prominence score. Comparing this surface with the one describing the interaction of syllable duration features with the corresponding features of the preceding syllable we hypothesise that the way neighbouring syllables influence prominence scoring is not symmetrical. The way nuclear durations interact, in this case, is clearly different than the way syllable durations do. As expected, Figure 10 shows that increasing Nucleus Duration matched with a short Nucleus Duration for the following syllables corresponds to increased prominence scoring. Nevertheless, a higher prominence score

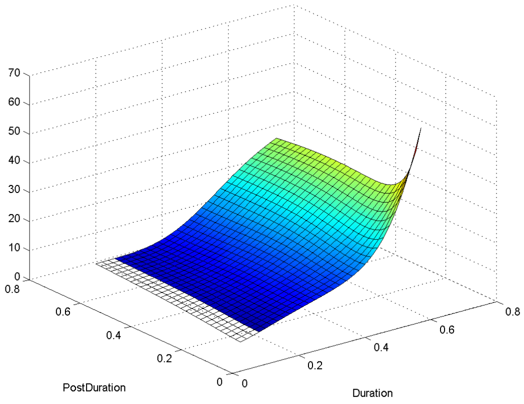

Fig. 9. Control surface for the interaction between the Duration of the following syllable and the Duration of the current syllable

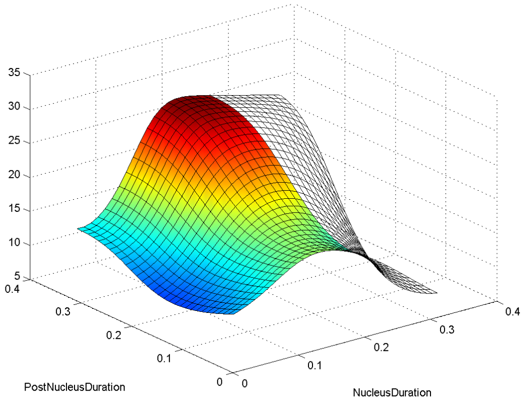

Fig. 10. Control surface for the interaction between the Nucleus Duration of the following syllable and the Nucleus Duration of the current syllable

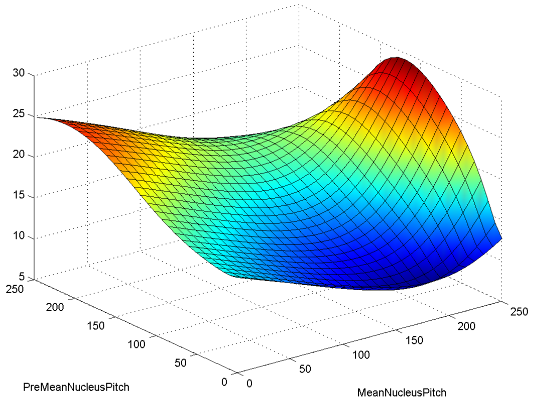

Fig. 11. Control surface for the interaction between the Mean Pitch of the preceding syllable and the Mean Pitch of the current syllable

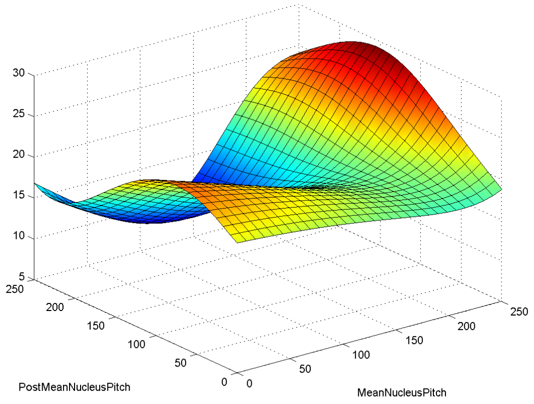

Fig. 12. Control surface for the interaction between the Mean Pitch of the following syllable and the Mean Pitch of the current syllable

is assigned to the syllable of interest if the following nucleus is long. This strengthening of the following nucleus on the considered syllable may correspond to some kind of phrasal interpretation.

Interactions between pitch features are summarised by the control surfaces in Figure 11 and in Figure 12. The obtained surfaces indicate that strong differences between average pitch values measured in the syllable nucleus yield higher prominence independently from the sign of the difference. Further investigation is needed on this specific comparison as the influence of high pitch matched with low pitch on the preceding syllable is weaker than expected. This may depend by the fact that, in this first set of experiments, we are not considering pitch dynamics inside the syllable nucleus, which are generally more important that simple average but tests with more complex features are needed to check this.

B. Quantitative analysis

In order to show that the application of ANFIS produces a TSK system able to identify the perceptual prominence in an opportune way, we perform a study about the performance of this system when it is used to predict the prominence in the dataset Speaker2. As for performance metrics, since we are dealing with a regression problem, we consider the well-known Root-Mean-Square Error $RMSE$. This metric is a frequently used measure of the differences between values predicted by a model and the values actually observed. The formal definition is as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \qquad (1)$$

where y is the vector of the observed values, $\hat{y}$ is the vector of the predicted values and n is the length of the vector y. In order to compute an evaluation independent from the unit/scale of the output variable, we consider together with RMSE other two evaluation metrics: Normalized Root-Mean-Square Error $NRMSE$ and the Coefficient of Variation of the RMSE CV_{RMSE}. Formally,

$$NRMSE = \frac{RMSE}{y_{max} - y_{min}} \qquad (2)$$

where y_{max} and y_{min} are respectively the maximum and the minimum of the vector of the observed values.

$$CV(RMSE) = \frac{RMSE}{\mu(y)} \quad (3)$$

where $\mu(y)$ is the mean of the observed values.

Table I shows the performance of the trained TSK system in terms of $RMSE$, $NRMSE$ and CV_{RMSE} by considering the dataset Speaker2.

TABLE I
PERFORMANCE OF THE TRAINED TSK SYSTEM ON DATASET SPEAKER2

$RMSE$	$NRMSE$	CV_{RMSE}
8.08	0.27	0.63

By analysing Table I, it is possible to put in evidence that the trained TSK is characterised by a good performance in terms of the capability of identifying the prominence when it is used on a dataset on which it has been not trained. These results validate the qualitative analysis in the previous section because this has been carried out on a good TSK model for prominence study.

V. CONCLUSIONS

The use of fuzzy systems, and in particular of the ANFIS inference system, has interesting applications in the domain of phonetic research. While most of the machine learning approaches used to investigate perceptual phenomena are designed with the specific goal of automatic annotation, advanced statistical modelling providing an interpretable description of the decision process estimated from training data may represent a powerful tool to investigate complex relationships among acoustic features. In the case of syllabic prominence, we have shown that ANFIS control surfaces can be interpreted from a phonetic perspective to deepen the understanding researchers have about how this is conveyed by human speakers. While the dataset considered in this work is limited, results coherent with the literature and interesting new perspectives have been reported. Future work will consist of applying this tool to larger datasets, also taking into account more complex features, like pitch movements inside the syllable nucleus.

ACKNOWLEDGMENTS

Antonio Origlia's work is supported by Veneto Region and European Social Fund (grant C92C16000250006).

REFERENCES

[1] P. Wagner, A. Origlia, C. Avesani, G. Christodoulides, F. Cutugno, M. DImperio, D. E. Mancebo, B. G. Fivela, A. Lacheret, B. Ludusan *et al.*, "Different parts of the same elephant: A roadmap to disentangle and connect different perspectives on prosodic prominence," in *Proc. of the 18th International Congress of Phonetic Sciences*, 2015.

[2] J. Terken, "Fundamental frequency and perceived prominence of accented syllables," *The Journal of the Acoustical Society of America*, vol. 89, no. 4, pp. 1768–1776, 1991.

[3] P. Wagner, A. Cwiek, and B. Samlowski, "Beat it! gesture-based prominence annotation as a window to individual prosody processing strategies," *Tagungsband der 12. Tagung Phonetik und Phonologie im deutschsprachigen Raum*, 2016.

[4] A. Rosenberg, E. Cooper, R. Levitan, and J. Hirschberg, "Cross-language prominence detection," in *Speech Prosody*, 2012.

[5] A. Lacheret, M. Avanzi, and B. Victorri, "A corpus-based learning method for prominence detection in spontaneous speech," in *Speech Prosody*, 2010, pp. 20–30.

[6] B. Ludusan, A. Origlia, and F. Cutugno, "On the use of the rhythmogram for automatic syllabic prominence detection," in *Proc. of Interspeech*, 2011, p. 24132416.

[7] A. Suni, J. Šimko, D. Aalto, and M. Vainio, "Hierarchical representation and estimation of prosody using continuous wavelet transform," *Computer Speech & Language*, 2016.

[8] F. Cutugno, E. Leone, B. Ludusan, and A. Origlia, "Investigating syllabic prominence with conditional random fields and latent-dynamic conditional random fields." in *Proc. of Interspeech*, 2012, pp. 2402–2405.

[9] F. Tamburini, C. Bertini, and P. M. Bertinetto, "Prosodic prominence detection in italian continuous speech using probabilistic graphical models," in *Proc. of Speech Prosody*, 2014, pp. 285–289.

[10] B. Heuft, *Eine pominenzbasierte Methode zur Prosodieanalyse und -synthese*. Peter Lang, Frankfurt, 1999.

[11] G. Fant and A. Kruckenberg, "Preliminaries to the study of swedish prose reading and reading style," *STL-QPSR*, vol. 2, no. 1989, pp. 1–83, 1989.

[12] O. Jespersen, *Lehrbuch der Phonetic*. B.G. Teubner, Leipzig e Berlin, 1920.

[13] A. Rietveld and C. Gussenhoven, "On the relation between pitch excursion size and prominence," *J. Phonet*, vol. 13, pp. 299–308, 1985.

[14] A. E. Turk and J. R. Sawusch, "The processing of duration and intensity cues to prominence," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3782–3790, 1996.

[15] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, pp. 341–345, 2001.

[16] G. Acampora and A. Vitiello, "Interoperable neuro-fuzzy services for emotion-aware ambient intelligence," *Neurocomputing*, vol. 122, pp. 3 – 12, 2013.

[17] N. B. Khameneh, H. Arabalibeik, P. Salehian, and S. Setayeshi, "Abnormal red blood cells detection using adaptive neuro-fuzzy system." in *MMVR*, 2012, pp. 30–34.

[18] H. Fang, "Adaptive neurofuzzy inference system in the application of the financial crisis forecast," *International Journal of Innovation, Management and Technology*, vol. 3, no. 3, p. 250, 2012.

[19] S. Kurnaz, O. Cetin, and O. Kaynak, "Adaptive neuro-fuzzy inference system based autonomous flight control of unmanned air vehicles," *Expert Systems with Applications*, vol. 37, no. 2, pp. 1229–1234, 2010.

[20] Y. Chakrapani and K. Soundararajan, "Adaptive neuro-fuzzy inference system based fractal image compression," *Dept. of Electronic and communications, JNTU College of Engineering, India*, vol. 2, no. 1, 2009.

[21] Y. Hamada, R. Elbarougy, and M. Akagi, "A method for emotional speech synthesis based on the position of emotional state in valence-activation space," in *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*. IEEE, 2014, pp. 1–7.

[22] M. E. Ekpenyong, U. G. Inyang, and E. O. Udoh, *Adaptive Prosody Modelling for Improved Synthetic Speech Quality*. Cham: Springer International Publishing, 2016, pp. 16–28.

[23] B. Pandey, A. Ranjan, R. Kumar, and A. Shukla, "Multilingual speaker recognition using anfis," in *2010 2nd International Conference on Signal Processing Systems*, vol. 3, July 2010, pp. V3–714–V3–718.

[24] S. L. Chiu, "Fuzzy model identification based on cluster estimation," *Journal of Intelligent & fuzzy systems*, vol. 2, no. 3, pp. 267–278, 1994.

[25] D. House, "Differential perception of tonal contours through the syllable," in *Proc. of ICSLP*, 1996, pp. 2048–2051.

[26] A. Origlia, G. Abete, and F. Cutugno, "A dynamic tonal perception model for optimal pitch stylization," *Computer Speech and Language*, vol. 27, pp. 190–208, 2013.