Corrective Focus Detection in Italian Speech Using Neural Networks

Asier López-Zorrilla[1], Mikel deVelasco-Vázquez[1], Sonia Cenceschi[2], M. Inés Torres[1]

[1] Speech Interactive Research Group, Universidad del País Vasco UPV/EHU
Barrio Sarriena s/n, 48940, Leioa, Spain
asier.lopezz@ehu.eus, mikel.develasco@ehu.eus, manes.torres@ehu.eus

[2] ARCSLab, Dep. of Electronics, Information and Bioengineering, Politecnico di Milano. Piazza Leonardo da Vinci 32, 20133, Milan, Italy
sonia.cenceschi@polimit.it

Abstract: The corrective focus is a particular kind of prosodic prominence where the speaker is intended to correct or to emphasize a concept. This work develops an Artificial Cognitive System (ACS) based on Recurrent Neural Networks that analyzes suitable features of the audio channel in order to automatically identify the Corrective Focus on speech signals. Two different approaches to build the ACS have been developed. The first one addresses the detection of focused syllables within a given Intonational Unit whereas the second one identifies a whole IU as focused or not. The experimental evaluation over an Italian Corpus has shown the ability of the Artificial Cognitive System to identify the focus in the speaker IUs. This ability can lead to further important improvements in human-machine communication. The addressed problem is a good example of synergies between Humans and Artificial Cognitive Systems.

Keywords: Focus; Stress; Prosodic prominence; Neural networks

1 Introduction

The stress prominence in speech is a phenomenon clearly related to human communication. Speakers usually focus acoustically one or more syllables of their speech in order to express emotions, which allows to position this work in the field of Affective Computing [1], or to introduce a new topic/concept into the dialog. Corrective focus is a particular kind of prosodic prominence where the speaker is intending to correct or to emphasize a concept. Thus, hereinafter we will refer to *focus* instead of citing the more general concept of prominence. The focus is a clearly cultural phenomenon, which is very dependent of the language and additional cultural facts. Thus, it is more frequent in some languages such as

English and Italian than in Spanish or French, that are very strong syllable-timed languages. The focus fits into the list of paralinguistic [2] and suprasegmental characteristics of human speech defined as prosody, involved in the cognitive processes of communicating and understanding. As a consequence, the automatic recognition of the occurrence of a prosodic prominence [3], or a focus in particular, in human speech is interesting for many different fields of study, Linguistics, Cognitive Sciences, etc. Moreover, it takes an important role in Human-Machine Communication.

In summary, the problem addressed in this work is the analysis of the intra-cognitive communication [4, 5] between a set of speakers who emphasized a word according to their communicative intention and a set of listeners aimed at detecting the focus in order to properly decode the message emitted by the sender. In this framework this work develops an Artificial Cognitive System (ACS) that plays the role of the listener resulting in inter-cognitive infocommunications [4, 5] between each speaker and the artificial system, thus using just the audio as the only CogInfoCom channel [6]. The ACS is based on Recurrent Neural Networks (RNNs) that analyzed suitable features of the audio channel. The capacities of such an artificial system are compared to the ones of the humans listeners allowing to analyze the synergies between Humans and artificial cognitive systems, i.e. between Engineering and Cognitive Sciences [7]. The results of our experiments showed the ability of the artificial cognitive system to identify the focus in the speaker IUs, which can result in further important improvements in human-machine communication [8].

The main novelty of this work lies in addressing the automatic focus detection with RNNs. This choice is based on the concept that the human speech is a continuous signal in the temporal domain where each syllable (focused or not) keeps a clear relation with the previous and following ones. In particular, we propose two different approaches to build the ACS. The first one is aimed at detecting focused syllables within a given utterance or Intonational Unit (IU), as explained in [9]. The second one identifies a whole IU as focused or not, so each of them address a different goal. Additional contributions refer to the proposed network structures that are powered only by the acoustic part of the message. Hence, the textual input is not required and as a consequence many technical problems can be bypassed allowing the methodology be improved and adapted to deal with other languages.

The experimental evaluation of the proposal was carried out over a subset of Italian Intonational Units based on the CALLIOPE Corpus [10, 11]. This corpus aims at cataloging IUs from an acoustic point of view, which agrees with our goal to investigate the prosody. Thus, we go beyond the analyses based on linguistic and language related contents, and consider the speech from a phonological and psychoacoustic point of view, as proposed in [12].

Section 2 deals with the pragmatic role and automatic detection of the corrective focus and includes some related works. Section 3 describes the two proposed approaches for the automatic corrective focus detection that are intended to reproduce the mechanism of understanding the focus normally unconsciously implemented at the cognitive level. Experiments carried out are fully described in Section 4. Section 4.2 shows the experiments carried out under the syllable-based approach whereas Section 4.3 deals with the experiments achieved at IU level. Section 4.4 includes a perceptual test concerning the focus recognition by Italian native speakers, allowing a comparison between the prediction ability of humans and ACSs. Finally, some concluding remarks are reported in Section 5.

2 Related Work

The stress prominence in speech [13] is a phenomenon that is easily and naturally produced and perceived by humans during a conversation. It is mainly produced with communicative purpose, but it is also related to the emotional status. Among the different kind of stress prominences, the corrective focus [14] is the main subject of research in this work. It consists in an acoustic stress applied to a syllable or entire word, in order to correct a content or a concept cited by the previous speaker.

Prosodic and paralinguistic cues have been largely explored in Natural Language Processing (NLP) [4], as well as the particular topic of the automatic detection of prominence [15]. Although textual information has been used in addition to acoustic features for the automatic focus detection [16], we are interested in working only with acoustic features because it simplifies the ACS and also makes it more language-independent. In this framework, [17] proposes a free-of-text automatic detection of stress on the Hungarian language at syllable level based on peaks of prosodic features.

If we consider Neural Networks methodologies in this area, the number of researches decrease considerably, and it is really limited narrowing down to the Italian language [18, 19]. Multiple types of stresses have been studied and classified with standard Feedforward Neural Networks [20, 21, 22] and with Convolutional Neural Networks [23, 24] with more success than other machine learning techniques. However, to the best of our knowledge RNNs have never been applied to detect the focus yet.

Another topic of interest regards the acoustic feature selection involved in focus characterization. Several studies have been carried out to determine which features are the most informative [15, 25, 26]. These seem to converge on variants of the same features: the duration of the focused syllable, the energy, the fundamental frequency contour, and the spectral emphasis. We report our own conclusions throughout the Section 4, where we show that the optimal feature selection depends on how the focus detection problem is addressed.

3 Automatic Corrective Focus Detection

The automatic recognition of the focus occurrence has a direct application for forensic or NLP purposes, where there is a need to identify new topics as well as a pragmatic and emotional discontinuities of the speaker on large amount of data. In such a case a procedure that works well at sentence level is needed. Distinctively, linguistic and phonology subjects, such as the characterization of dialects or the learning of a language, might require a more refined system allowing to get the time position of the focus into a word.

As a consequence we propose to formalize two different pattern recognition tasks to be solved. In the first task a given syllable in an IU has to be classified as *focused* or *not focused*. To this end the acoustic features of the given syllable as well as its previous and following temporal context will be considered. This task was named as the *focus in syllables classification problem* (FSP).

The second task will deal with whole IUs. In this case, the ACS will predict if any syllable in the sentence has been uttered with a corrective focused or not. Therefore, the acoustic features will be calculated at regular time windows in the whole IU. This task will be referred as the *focus in IUs classification problem* (FIUP).

3.1 The Focus in Syllables Classification Problem (FSP)

This Section describes the FSP approach aimed at detecting focused syllables in given IUs. The section first includes some details of the feature extraction methodology for this problem, then it explains two ways to combine these features in order to build the input of the classifiers, and it finally describes the structure of the proposed Neural Networks.

Feature extraction. The feature extraction procedure was based on a short-term analysis of the speech signal over 25 ms windows overlapping each 10 ms. For each frame we extracted: Pitch, Zero-Crossing Rate (ZCR), Energy, the Spectral Centroid, Spectral Spread, 13 Mel-frequency Cepstral Coefficients (MFCCs), 16 Linear Predictive Coefficients (LPCs) and 29 Bark features[1]. Additionally, we

[1] Pitch, LPCs, and Bark features were extracted with the Praat Speech Analysis Tool [27] whereas ZCR, energy, spectral centroid, spectral spread and MFCCs with the PyAudioAnalysis library.

computed the first and second derivatives of these 63 features, which increased the number of available features per frame to 189.

Then this number was increased again to 378 by adding the long-term smoothed features of the short-term ones. The smoothing was carried out by calculating the average value of the short-term features centered on the given frame. The number of feature vectors involved in that average were 23 (the central one and 11 previous and following vectors). This time interval is very close to the mean syllable duration in Italian: (0.235 ± 0.1) s[2]. This makes sense because the problem to be solved is the detection of focused syllables which are quasi-stable during their duration. Finally, every feature vector was normalized so that its mean and standard deviation per IU are 0 and 1 respectively.

Building the input to the classifiers. In order to build the input vector to be supplied to the classifier we assume that each IU in the corpus is segmented into syllables, i.e. that we know when each syllable starts and ends. Thus, given a syllable in a IU the input vector will consist of the feature vector corresponding to the center of the syllable under consideration along with some additional feature vectors representing the syllable context as well as the duration (in seconds) of the syllable. At this point two different methods to get such a context were proposed: a fixed frame distance and a context size related to the syllable duration.

> **Fixed frame distance.** In this approach both the context size and the frame distance are fixed. The first refers to the number of left and right context feature vectors that will be selected, whereas the second to the distance between consecutive context vectors. As an example, if the context size is fixed to 2 and the frame distance equals 3, the input would be built as in the Figure 1.

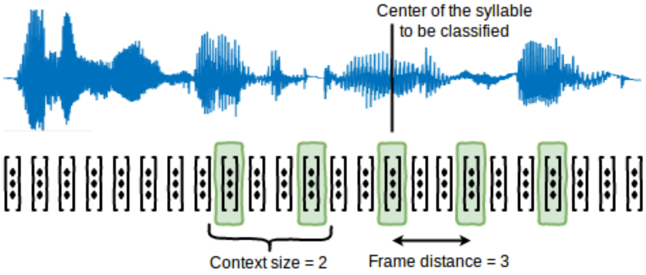

Figure 1
An example of how to build the input with fixed frame distance. In this case, 5 vectors were taken in total: the central one and two left and right context vectors, according to the context size. The frame distance was set to 3.

[2] This value was computed after an automatic syllable segmentation process of our corpus.

Beginning, center and ending of neighbor syllables. In this approach the context feature vectors are selected among the ones representing the beginning, the center and the end of the neighbor syllables, according to the segmentation of the IU into syllables. Hence, in this case we only need to specify the context size. Figure 2 shows an example of the input for a context size set to 3.

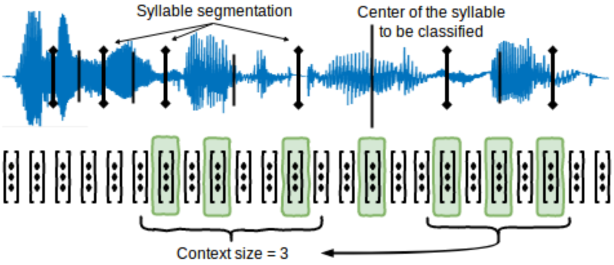

Figure 2
An input built using feature vectors corresponding to the beginning, center and ending of neighbor syllables. Since the context size was set to 3, the vectors corresponding to the end of the central syllable (which is also the one that corresponds to the beginning of the next syllable), to the center of the next syllable and to the end of the next syllable were selected as the right context. Symmetrically, the left context consists of the vectors corresponding to the beginning of the central syllable, to the center of the previous syllable and to the beginning of it.

Classifiers. The previous methods allow the generation of training examples that can be used by common machine learning algorithms. Once the specific set of acoustic features are selected and the methodology to build the input is chosen, all the feature vectors can be concatenated to form a fixed-dimensional input vector representing each syllable in the corpus. Then, classifiers such us Naive Bayes, Support Vector Machines (SVM) and conventional Feedforward Deep Neural Networks can be directly trained. These classifiers were used for the experiments shown in Section 4.2. However, the temporal relationship between the feature vectors that compose the input of each training example is not considered enough by these classifiers. Thus, more complex neural networks based on recurrent layers might be more suitable. In this framework we propose RNNs with two parallel sets of recurrent layers. The first one processes the left (previous in time) context vectors forward, i.e., it takes first the farthest context vector in the left-side and sequentially all the left context vectors until the central vector is processed. Symmetrically, the other set of recurrent layers processes the right context vectors backwards. Additionally, our architecture includes another parallel set of feedforward layers, which processes the scalar corresponding to the duration of the syllable we want to classify. Finally, the three sets are merged and the network ends with a set of feedforward layers. Figure 3 shows a graphical representation of the proposed Bidirectional RNNs. These networks led to the best system performance when dealing with the FSP according to the experiments carried out (see Section 4.2).

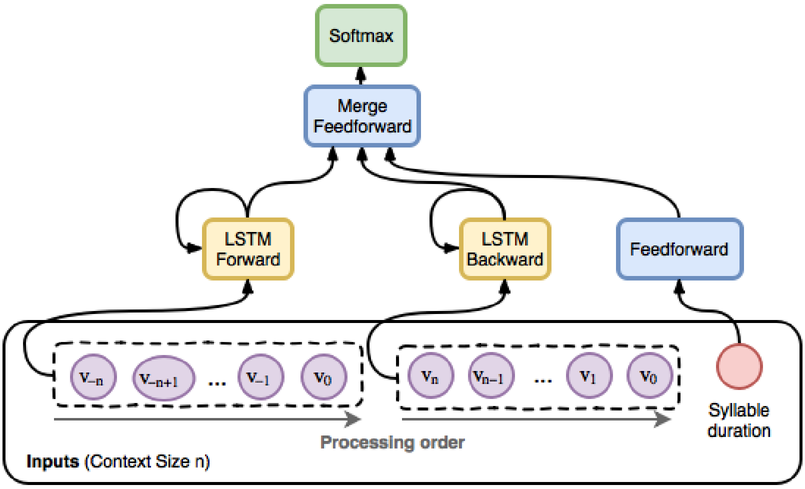

Figure 3
An example architecture of a neural network used in the FSP. The two sets of recurrent layers consist of a single LSTM layer each. The duration of the syllable is also processed with a single feedforward layer. Then the output of these three layers are merged into a feedforward layer followed by a softmax layer of two outputs, one per class.

3.1 The Focus in IUs Classification Problem (FIUP)

This Section describes the FIUP approach aimed at classifying a whole IU as focused or not. The feature extraction methodology for this problem is the one used to deal with FSP problem. Thus, this section just explains the way to combine these features in order to build the input of the classifiers, and then it describes the structure of the proposed Neural Networks.

Building the input to the classifiers. We propose two different ways to build the input of the classifiers: the first one is based on regular sampling of the sequence of feature vectors whereas the second one is based on the output of the networks classifying syllables (FSP) as focused or not focused.

> **Fixed frame distance.** If we use a fixed sampling rate from the beginning to the end of the IU to select the feature vectors that will be involved in the classification process, more than one training example per IU can be generated. More precisely, if the frame distance was set to *n*, we can generate *n* examples, just alternating the vector from where the sampling starts.
>
> **From the FSP to the FIUP.** In this approach we take advantage of the classifiers trained to solve the FSP. Each given IU can be automatically segmented into (pseudo-)syllables. Then, the input corresponding to each of

these pseudo-syllables can be propagated across an already trained classifier. Afterwards, these predictions can be used as an alternative input to train a classifier to deal with FIUP. This approach is specially interesting if the classifier trained to solve the FSP is a Neural Network, since not only its output can be used, but also the output of the penultimate layer, which contents more features about the syllable.

Classifiers. An additional difference between the FSP and the FIUP approaches is that common classifiers cannot directly be trained. In fact, Naive Bayes and SVMs classifiers as well as Feedforward Neural Networks require the dimension of input vector to be fixed for all the examples. However, such a condition will certainly not be met due to the variable length of the IUs (if we are using the first way to build input), and/or because of the variable number of syllables in the IUs.

RNNs, though, are still directly trainable in this scenario. These are able to sequentially process any sequence of vectors of arbitrary length, which makes them really suitable for this task. In particular, we propose bidirectional RNNs. One set of layers processes the whole sequence of feature vectors forwards, from the first vector to the last. Another set of layers processes the sequence in the inverse order, backwards from the last vector to the first. Figure 4 shows a graphical representation of the proposed structure. Note that the proposed RNN is able to deal with inputs obtained under the two building methodologies proposed.

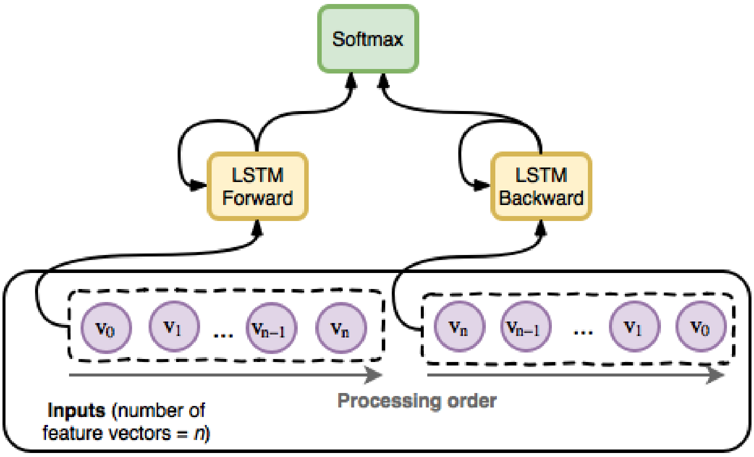
Softmax
LSTM Forward
LSTM Backward
v_0 v_1 ... v_{n-1} v_n
v_n v_{n-1} ... v_1 v_0
Processing order
Inputs (number of feature vectors = *n*)

Figure 4
An example RNN used in the FIUP. A LSTM layer processes the input forwards and another forward. The network ends in a softmax layer of two outputs.

4 Experimental Study

Two series of experiments were carried out to evaluate the performance of the ACS. The first series aims to validate the proposals described in Section 3.1 when dealing with the FSP whereas the second one focus in the FIUP under the approaches proposed in Section 3.2. An additional set of experiments allow to analyse the human perception abilities for the same data collection. A subset of the Standard Italian Corpus (SIC) described in Section 4.1 was used for all the experiments.

4.1 The Standard Italian Corpus

Italian is a romance, iso-syllabic and free-stress language [19]. Then, the position of a contrastive focus is just a communicative choice of the speaker. The presence of focus has been related to the duration of the syllable, or to the distance between peaks of energy (syllable nuclei). In fact, the duration of a focused syllable is typically higher than the one of not focused syllables of the same speaker. However, it is unrelated to the tonic/tonic syllables alternations providing the rhythm [26].

The Corpus selected to carry out the proposed series of experiments is based on CALLIOPE (Combined and Assessed List of Latent Influences On Prosodic Expressivity), a conceptual model created within the LYV project[3] aiming at categorizing all IUs. Each IU is thus associated to a "point" into this space and associated to a tuple composed of 12 labels (detailed descriptions in [10]). In this multidimensional space each dimension represents a characteristic influencing the vocal paralinguistic components of the speech assuming values in a set of labels.

Table 1

List of the CALLIOPE dimensions

Group	**Dimensions (F_i)**
Dialogic	Structure (F_1), Linguitic modality (F_2), Intonational focus (F_3), Rhetorical form (F_4), Motivational state (F_5), Speech mood (F_6), Spontaneity (F_7), Punctuation forms (F_8), Emotions (F_9)
Background	Expressiveness skill (F_{10}), Social context (F_{11}), Launguage (F_{12})

Each IU has a subjective correspondence with a specific prosodic unit. Starting from this conceptual model a database of Italian standard speech has been defined and created. CALLIOPE dimensions are divided into two groups as shown in Table 4.1. The Dialogic group contains characteristics directly related with the

[3] LYV is a project of the Polisocial program 2016-2017, http://www.polisocial.polimi.it focused on the improvement of prosodic and expressive skills of Italian speakers with cognitive disabilities, through the use of technology in complex contexts [28].

communication context, where the corresponding sets of labels are fully defined. The second group contains background dimensions, i.e, characteristics that exist regardless of whether or not there is an interaction.

The selected corpus concerns a subspace of the CALLIOPE model, obtained narrowing the field of recordings by setting 6 dimensions as follows. The language (F_{12}) is the Standard Italian [29], recited by able-bodied (F_{10}) actors (F_7) and the contents concern daily situations (F_{11}) and absence of particular motivational states (F_5) emotions (F_9). The corpus considers 13 Calliope's labels (among the remaining 6 dimensions) an includes the Corrective Focus, which was validated by a perceptive test performed on about 200 Italian native-speakers. Audio files were recorded in WAV format (44.1 kHz 16 bit) with different modes and microphones to obtain a model as independent as possible from the technical apparatus. 14 speakers (7 men and 7 women) aged between 33 and 48 were recorded. Each speaker recorded 278 IUs (139 with meaning and 139 pseudo-sentences [30] with equal prosody) so that the corpus contains 1946 sentences with meaning and 1946 pseudo-sentences. Considering both real and pseudo sentences, 2884 IUs do not contain any prosodic prominence while 1008 contains one or more corrective focuses.

This database is ready for the experimental evaluation of the proposals to solve the FIUP through the second series of experiments. However, the FSP needs a segmentation of each IU into syllables that have to be labelled. To this end we proposed an automatic syllable segmentation procedure that was based on the syllable positions provided by Praat [27], i.e. the beginning and end of each syllable. Some few errors appeared for long syllables that were sometimes split into two subsegments. Then, we manually labeled each of these (pseudo-)syllables as *focused* or *not focused*. In total, the resulting corpus consists of 44923 pseudo-syllables; 1867 focused and 43056 not focused. This corpus is highly unbalanced and includes one focused pseudo-syllable per 22 non focused ones, approximately.

4.2 Study of the FSP

Preliminar experiments. The initial experiments included the parametric Naïve Bayes classifier and the geometric SVM one as well as Feedforward Neural Networks. The average F1-score between the two classes in our dataset was used to evaluate the performance of each classifier. This measure was computed after a 7-fold cross-validation process. In each iteration the instances of 2 of the 14 speakers in the corpus were left as the test partition. All the neural networks were implemented with the WBNN toolkit[4], while the Scikit-learn toolkit was chosen to

[4] The first and second authors of this work are the main developers of this open source toolkit,. which is still under development. It can be found at https://github.com/develask/White-Box-Neural-Networks.

train the Naïve Bayes and the SVM classifiers. Columns 2 to 4 in Table 2 show the results of these experiments and confirm that Neural Networks outperform both SVM and Naive Bayes classifiers in terms of the average F1-score.

Table 2

Average F1-score obtained by different classifiers

	Best RNN	**Best feedforward NN**	**Best SVM**	**Best Naïve Bayes**
Average F1-score	0.693	0.618	0.576	0.512

Experiments with the proposed Recurrent Neural Networks. We then focused on bidirectional RNNs due to their ability to process sequences of variable length. In particular, we explored several RNN architectures and hyperparameters as well as several ways to build the input to the network and its parameters. First column, in Table 2 shows the best results that were achieved with RNN that clearly outperfomed the ones obtained by Feedforward NN. The structure of this best RNN is very similar to the one previously shown in Figure 3. Each recurrent layer consists of 10 LSTM cells[5], the layer that processes the syllable duration is made of 8 sigmoidal units, the layer after merging the three sets of parallel layers consists of 20 sigmoidal units, and the network ends in a softmax layer of two units, one per class. Results in column one in Table 2 were obtained when the set B of features (pitch, energy and spectral centroid without any derivative) was selected. Finally, a fixed frame distance of 11 and a context size of 9 vectors resulted to be the best configuration to build the RNN input. The RNNs were trained by stochastic gradient descent with an exponentially decaying learning rate during a fixed number of epochs. The best choice for theses parameters was to reduce the learning rate from 0.5 to 0.1 throughout 75 epochs.

This is the configuration for the ACS achieving the higher system performance shown in Table 2, i.e. the best RNN. To get these results we had previously explored two techniques to deal with the imbalance of the data set. We first included a classical variable decision threshold to determine the confidence level[6] required by the RNN to predict that the input corresponds to a focused syllable. An exhaustive search of this parameter was carried out to maximize the average F1-score between the two classes in the training partition. As an alternative we proposed to apply an increasing imbalance schedule in the training data [32]. To this end the network was trained with different data each epoch, starting from a not very unbalanced subset of the training data and slowly adding more examples from the majority class. The best schedule was to increase the imbalance from 5 (5 non-focused syllables per each focused one) to the real imbalance (around 22),

[5] We implemented the LSTM version proposed in [31].

[6] The confidence level is the output of the neuron of the softmax layer that corresponds to focused syllables.

with a scaled hyperbolic tangent function. Table 3 shows how the performance was improved with the use of these techniques.

Table 3
Average F1-score obtained with the proposed techniques to deal with unbalanced data

	RNN with threshold and imbalance schedule	**RNN with threshold**	**Baseline RNN**
Average F1-score	0.693	0.618	0.576

Effect of the sets of features. We explored a variety of features as well as several ways to combine them. Then, the six sets of features listed below were selected. Additionally, we also experimented with sets that added the first derivatives of the proposed features on the one hand or the first and the second derivatives on the other hand. Note that all the features correspond to the long-term smoothed version.

Set A. Pitch and energy.
Set B. Pitch, energy and spectral centroid.
Set C. Pitch, energy, spectral centroid, ZCR and spectral spread.
Set D. Pitch, energy, spectral centroid, ZCR, spectral spread and 13 MFCCs.
Set E. Pitch, energy, spectral centroid, ZCR, spectral spread and 16 LPCs.
Set F. Pitch, energy, spectral centroid, ZCR, spectral spread and 29 Bark features.

Figure 5 shows the performance of the described best model when different sets of features were used. First and second derivatives led to a decrease of performance for all the feature sets. i.e. they did not add any information. Pitch, energy and spectral centroid resulted to be the most informative features for this problem. The high performance obtained by the ACS when a so reduced set of features was used outlines the capability of the proposed RNN structure and configuration.

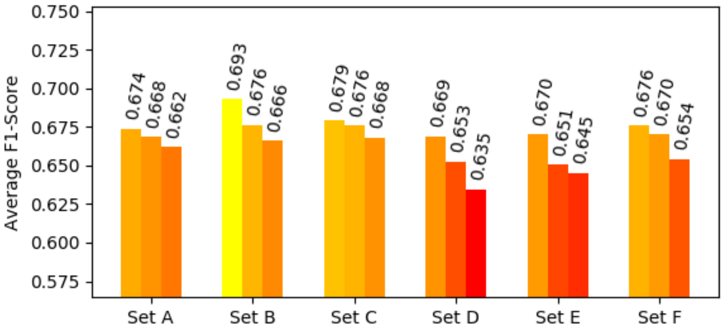

Figure 5
Average F1-score obtained with the best network trained with different sets of features. The three columns showed per set indicate the performance when no derivatives are added (left column), when the first derivative is added (central column) and when the first and second derivatives are added (right column).

Effect of the context. Figure 6 shows the ACS performance of the described best model and best set of features for different values of the context size and frame distance as defined in Section 3.1. Figure 6 evidences that a lack of information, i.e. a small context size, drastically worsens the system's performance. However, big context sizes do not significantly reduce the classification capacity of the proposed ACS. Thus, the ability of the LSTMs to *forget* non relevant events appear to pay off but the computation time is clearly much higher. On the other hand, the analysis of the frame distance shows an optimal range between 5 and 15 frame distance where the performance does not significantly depend on the value of this parameter. However, the average F1-score clearly decays out of this range. Thus, low frame distances considers a few context but very big ones seem to lead to a loss of important events.

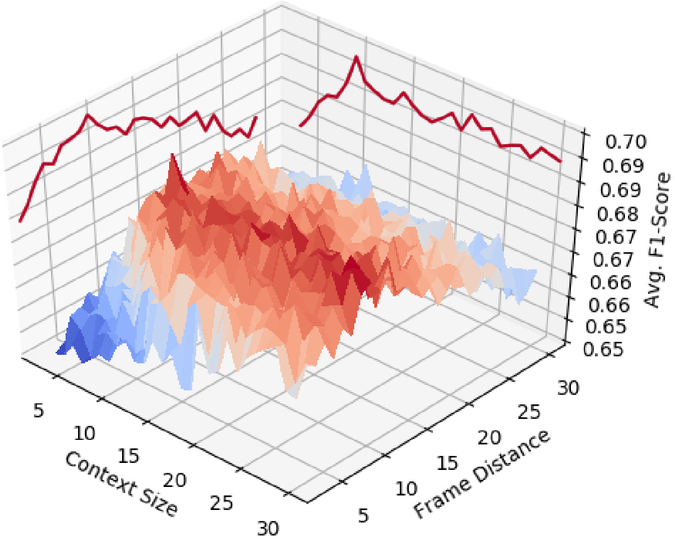

Figure 6

Average F1 score in the FSP of the described best network and best set of features for different values of the context size and frame distance as defined in Section 3.1

4.3 Study of the FIUP

A second series of experiments were carried out with the Standard Italian Corpus in order to deal with the FIUP. The sets of parameters defined in Section 4.2 were also considered for these experiments.

Experiments with the proposed RNNs. The RNNs proposed to solve the FIUP are based on the architectures described in Figure 4. The best results were obtained when 60 LSTM cells per recurrent layer were considered and the RNN was trained during 40 epochs. The best learning rate schedule was still an exponentially decaying one from 0.5 to 0.1. In addition, a variable decision threshold was included to optimize the average F1-score in the training partition. However, the use of a schedule throughout the epochs to deal to the imbalance at training time did not lead to any improvement in this case. This is probably due to the fact that the imbalance is not so high in the FIUP (around 3 IUs without focus per each IUs with focus).

Effect of the sets of features. Figure 7 shows the performance of the described best RNN when different sets of features were used. Unlike the FSP problem the first derivatives seem to be significant mainly for set F. In fact, the size window analysis is now bigger so that the information provided by derivatives is meaningful. Moreover, Set F, which consists of the pitch, the energy, the spectral centroid, ZCR, the spectral spread and 29 Bark features, led to the higher ACL performance for this problem achieving a great average F1-score of 0.826. In the same way spectral changes seem also to be more significant for larger windows.

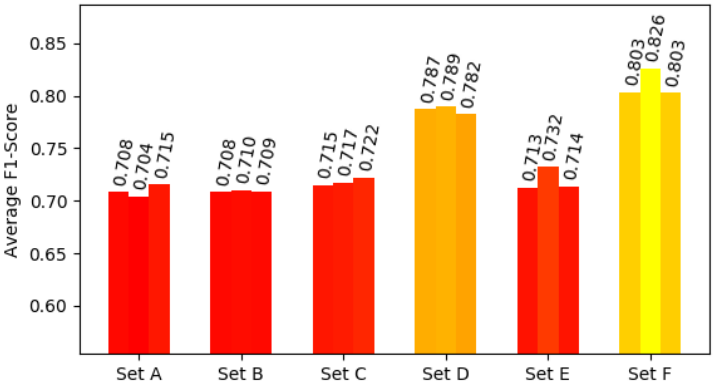

Figure 7

Average F1-score obtained with the best network trained with different sets of features for the FIUP. As before, the columns represent the addition of no derivatives, the addition of the first derivatives, and the addition of the first and second derivatives.

Effect of the context. When dealing with the FIUP the context is just represented by the frame distance at which input vectors at subsampled. Figure 8 shows the ACS performance of the described best model and best set of features for different values of the frame distance as defined in Section 3.2. Figure 8 evidences a similar effect of the frame distance in system performance than the one analyzed for FSP. In fact, Figure 8 still shows an optimal range where the performance does not significantly depend on the value of this parameter and a very strong decrease of F1-score out of this range. Thus, once again big frame distances seem to lead to a loss of important information.

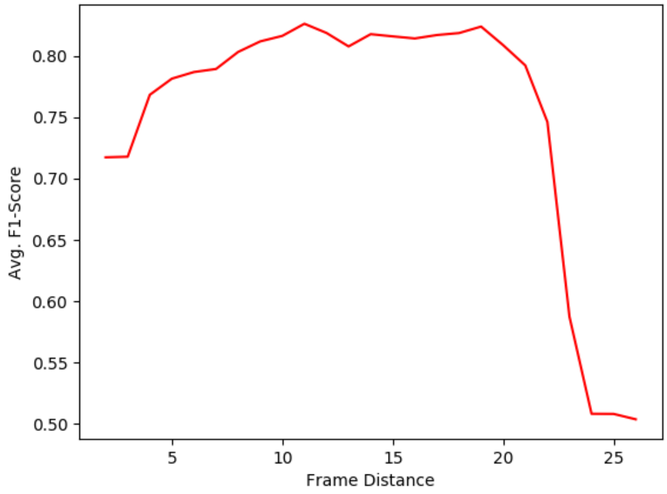

Figure 8

Average F1 score of the described best network and best set of features for different values of the frame distance in the FIUP as defined in Section 3.2

From the FSP to the FIUP. Figure 9 shows the results when predictions from previous FSP classifier were used as inputs for RNN proposed in Section 4.2 to deal with FIUP. Figure 9 evidences that the ACS performances are now lower than the ones got by the previous direct approach. However, let us note that the best result (a F1-score of 0.756) was obtained with an RNN trained on top of the outputs (of the last and penultimate layers) of a network that processes the Set F of features, with no derivatives. Thus, the spectral information seem to be also meaningful with this approach when dealing with the FIUP.

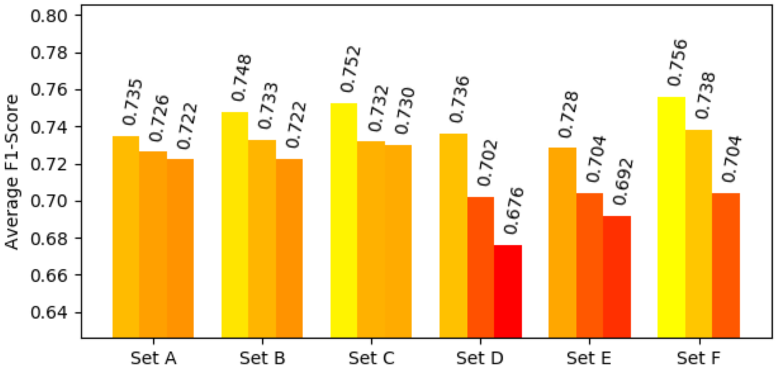

Figure 9

Average F1-score got when training RNNs on top of the features extracted with FSP classifiers

4.4 Human Perception Tests for the FIUP

A series of Human Perception Tests was also carried out with the Italian Corpus. To this end a set of 203 adults, Italian native-speakers, were asked to recognize the 13 labels mentioned in Section 4.1. They classified all the sentences and pseudo-sentences in the corpus without repetitions, i.e., only one speaker per IU. In this work we just considered the question related to the presence of a corrective focus. The average F1-score between the two classes was 0.444, which is much lower than the performance got by both approaches of the ACS dealing with the same FIUP, i.e. 0.826 and 0.756 respectively in terms of F1 scores.

The low perceptive rates may be due to the listener's need of the context provided by a previous interaction of other speaker. It seems that one single IU is not enough to ensure the human focus recognition. In contrast, with the narrow context preferred by the ACS, the human auditory apparatus seems to require a very broad one, extending to other parts of the dialogue.

Concluding Remarks

The corrective focus is a particular kind of prosodic prominence where the speaker is intended to correct or to emphasize a concept. This work has developed an Artificial Cognitive System (ACS) that played the role of the listener resulting in inter-cognitive infocommunication between a speaker and the artificial system, thus using just the audio as the only CogInfoCom channel. The ACS is based on Recurrent Neural Networks that analyze suitable features of the audio channel. Two different approaches to build the ACS has been developed. The first one addressed the detection of focused syllables within a given Intonational Unit whereas the second one identify a whole IU as focused or not. For the first problem the proposed RNN achieved an F-score of 0.693 with a reduced set of acoustic features whereas the RNN were able to get a really high F1-score of 0.826, with a larger set of acoustic features that also includes variations. Experimental results showed the need of context to detect the focus. However, this context is reduced to neighbor syllables. On the other hand, human perception experiments showed that Humans were able to get just an F1 score of 0.444 probably due to the lack of broad contexts including previous dialog turns.

The results of our experiments showed the ability of the Artificial Cognitive System to identify the focus in the speaker IUs, which can lead to further important improvements in human-machine communication. The behavior of the ACS to identifies the focus in speech that can be interpreted, to some extend, as an estimation, optimistic in this case, of the human cognitive load when dealing with the same problem, showing synergies between Humans and Artificial Cognitive Systems.

Acknowledgements

European Commission

The research leading to the results in this paper has been conducted in the project EMPATHIC (Grant N: 769872) that received funding from the European Union's Horizon 2020 research and innovation programme.

Additionally, this work has been partially funded by the Spanish Minister of Science under grants TIN2014-54288-C4-4-R and TIN2017-85854-C4-3-R, by the Basque Government under grant PRE_2017_1_0357, and by the University of the Basque Country UPV/EHU under grant PIF17/310.

References

[1] Lisetti, C. L. (1998) Affective Computing

[2] Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., & Narayanan, S. (2013) Paralinguistics in Speech and Language—State-of-the-Art and the Challenge. Computer Speech & Language, 27(1) 4-39

[3] Terken, J. (1991) Fundamental Frequency and Perceived Prominence of Accented Syllables. The Journal of the Acoustical Society of America, 89(4) 1768-1776

[4] Baranyi, P., & Csapó, Á. (2012) Definition and Synergies of Cognitive Infocommunications. Acta Polytechnica Hungarica, 9(1) 67-83

[5] Baranyi, P., Csapó, Á., & Sallai, G. (2015) Cognitive Infocommunications (CogInfoCom) Springer

[6] Fulop, I. M., Csapó, Á., & Baranyi, P. (2013, December) Construction of a CogInfoCom ontology. In Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on (pp. 811-816) IEEE

[7] Irastorza, J., & Torres, M. I. (2016, October) Analyzing the Expression of Annoyance during Phone Calls to Complaint Services. In Cognitive Infocommunications (CogInfoCom) 2016 7th IEEE International Conference on (pp. 000103-000106) IEEE

[8] Torok, A. (2016, October) From Human-Computer Interaction to Cognitive Infocommunications: A Cognitive Science Perspective. In Cognitive Infocommunications (CogInfoCom) 2016 7th IEEE International Conference on (pp. 000433-000438) IEEE

[9] Cresti, E. (2000) Spoken Italian Corpus: an Introduction [Corpus di italiano parlato: Introduzione] (Vol. 1) Accademia della Crusca

[10] Cenceschi, S., Sbattella, L., & Tedesco, R. (2018) Towards Automatic Recognition of Prosody. In Proceedings of 9th International Conference on Speech Prosody 2018 (pp. 319-323)

[11] Sbattella, L., Tedesco, R., & Cenceschi, S. (2017) The Definition of a Descriptive Space of Italian Prosodic Forms: The CALLIOPE Model. In XIII Convegno Nazionale AISV (pp. 1-3) ITA

[12] Dominguez, M., Farrús, M., & Wanner, L. (2016) An Automatic Prosody Tagger for Spontaneous Speech. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (pp. 377-386)

[13] Werner, S., & Keller, E. (1995, May) Prosodic Aspects of Speech. In Fundamentals of Speech Synthesis and Speech Recognition (pp. 23-40) John Wiley and Sons Ltd.

[14] Gussenhoven, C. (2008) Types of Focus in English. In Topic and focus (pp. 83-100). Springer, Dordrecht

[15] Tamburini, F. (2003) Automatic Prosodic Prominence Detection in Speech using Acoustic Features: an Unsupervised System. In Eighth European Conference on Speech Communication and Technology

[16] Beke, A., & Szaszák, G. (2014, November) Combining NLP Techniques and Acoustic Analysis for Semantic Focus Detection in Speech. In 5th IEEE Conference on Cognitive Infocommunications (CogInfoCom) 2014 (pp. 493-497) IEEE

[17] Tündik, M. Á., Gerazov, B., Gjoreski, A., & Szaszák, G. (2016, October) Atom Decomposition-based Stress Detection and Automatic Phrasing of Speech. In Cognitive Infocommunications (CogInfoCom) 2016 7th IEEE International Conference on (pp. 000025-000030) IEEE

[18] Tamburini, F., Bertini, C., & Bertinetto, P. M. (2014) Prosodic Prominence Detection in Italian Continuous Speech using Probabilistic Graphical Models. In Proceedings of Speech Prosody (pp. 285-289)

[19] Kori, S., Farnetani, E., & Cosi, P. (1987) A Perspective on Relevance and Application of Prosodic Information to Automatic Speech Recognition in Italian. In European Conference on Speech Technology

[20] Jenkin, K. L., & Scordilis, M. S. (1996, October) Development and comparison of three syllable stress classifiers. In Spoken Language, 1996 ICSLP 96 Proceedings, Fourth International Conference on (Vol. 2, pp. 733-736) IEEE

[21] Li, K., Qian, X., Kang, S., & Meng, H. (2013) Lexical Stress Detection for L2 English Speech Using Deep Belief Networks. In Interspeech (pp. 1811-1815)

[22] Shahin, M. A., Ahmed, B., & Ballard, K. J. (2014, December) Classification of Lexical Stress Patterns Using Deep Neural Network Architecture. In Spoken Language Technology Workshop (SLT) 2014 IEEE (pp. 478-482) IEEE

[23] Heba, A., Pellegrini, T., Jorquera, T., André-Obrecht, R., & Lorré, J. P. (2017, October) Lexical Emphasis Detection in Spoken French Using F-BANKs and Neural Networks. In International Conference on Statistical Language and Speech Processing (pp. 241-249) Springer, Cham

[24] Stehwien, S., & Vu, N. T. (2017) Prosodic Event Recognition using Convolutional Neural Networks with Context Information. arXiv preprint arXiv:1706.00741

[25] Streefkerk, B. M. (1997) Acoustical Correlates of Prominence: A Design for Research. In Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam (Vol. 21, pp. 131-142)

[26] Giordano, R. (2008, May) On the Phonetics of Rhythm of Italian: Patterns of Duration in Pre-planned and Spontaneous Speech. In Proceedings of the 4th Speech Prosody Conference, Campinas, BR

[27] Boersma, P. (2006) Praat: Doing Phonetics by Computer. http://www. praat. org/

[28] Sbattella, L. (2006) La mente orchestra. Elaborazione della risonanza e autismo. Vita e Pensiero

[29] Canepari, L. (1980) Italiano standard e pronunce regionali. Cooperativa libraria editrice degli studenti dell'università di Padova

[30] Cibelli, E. (2012) Shared Early Pathways of Word and Pseudoword Processing: Evidence from High-Density Electrocorticography

[31] Graves, A., & Schmidhuber, J. (2005) Framewise Phoneme Classification with Bidirectional LSTM and other Neural Network Architectures. Neural Networks, 18(5-6) 602-610

[32] López-Zorrilla, A., de Velasco-Vázquez, M., Serradilla-Casado, O., Roa-Barco, L., Graña, M., Chyzhyk, D., & Price, C. C. (2017, June) Brain White Matter Lesion Segmentation with 2D/3D CNN. In International Work-Conference on the Interplay Between Natural and Artificial Computation (pp. 394-403) Springer, Cham