9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

Towards Automatic Recognition of Prosody

Sonia Cenceschi, Licia Sbattella, Roberto Tedesco

Politecnico of Milano, Italy

sonia.cenceschi@polimi.it, licia.sbattella@polimi.it, roberto.tedesco@polimi.it

Abstract

The term *prosody* defines the group of audio paralinguistic and suprasegmental cues involved in the communicative and understanding process of human speech. This paper presents our approach to automatic recognition of prosodic forms. In particular, we present: CALLIOPE, a multi-dimensional and abstract model, aiming at categorising all prosodic forms; SI-CALLIOPE, a sub-space for which we defined a corpus of recorder prosodic forms; and the psychoacoustic experiment we are currently carrying on for investigating main acoustic behaviours and features involved into the discrimination of prosodic forms. The experiment results will be useful for defining the feature set to rely on for automatic recognition of prosodies. For that reason, we are also defining a classifier, based on Neural Nets. This study is part of the LYV project, which focuses on improving prosodic expressiveness skills of Italian speakers with autism and other cognitive disabilities.

Index Terms: prosody, perception, paralinguistic interaction, Neural Network.

1. Introduction

Prosody is composed by suprasegmental audio features, and their variations on time; in particular: rhythm, speech rate, pitch contour, intensity, and pauses. Prosodic forms are affected by several contextual factors, like speaker's language and culture, emotional state, presence of irony or sarcasm, etc [1]. All of such factors mix together and contribute to produce the final prosodic "sound".

To cope with this complexity, and recognise the right prosody, the human brain needs to mix acoustic information and previous knowledge [2]. Thus, our first step was defining a conceptual model that formalises such knowledge as a set of factors affecting prosody emission.

The result was the CALLIOPE (Combined and Assessed List of Latent Influences On Prosodic Expressivity) model [3]. CALLIOPE is a conceptual multidimensional space, defining all possible independent factors affecting prosody.

Then, as a second step, we needed to validate our model. Unfortunately, being CALLIOPE a very general model, it is too big to be validated as a whole. Thus, we decided to focus on a subspace, which we called SI-CALLIOPE (Standard Italian CALLIOPE), containing standard Italian prosodic forms. Such space was also useful for the goals of our LYV (Lend Your Voice) [4] which focuses on improving prosodic expressiveness of Italian speakers with autism and other cognitive disabilities [5].

For validating SI-CALLIOPE, we recorded a corpus of Standard Italian (as defined in [6]) prosodic forms, and used them to design a perceptive experiment with Italian native listeners. Comparing experiment results with analysis of audio signals, we also defined a list of relevant audio features involved into the psychoacoustic discrimination of prosody. Finally, we defined a Neural Network-based classifier for the automatic recognition of a subset of prosodic forms.

We think the CALLIOPE model, allowing a precise description of factors affecting prosody, could be helpful to guarantee repeatability and accuracy of experiments, in different fields of study; this as highly desirable as, to our knowledge, no models exist that try to provide a multidimensional classification of prosodies. So far, many studies focus on detecting single prosodic cues [7] using different classifiers; for example the stress prominence [8, 9].

2. The CALLIOPE model

CALLIOPE is a model aiming at categorising all prosodic forms. We will call sentences as Information Units (IU) [10]. Each IU has a bi-univocal correspondence with a specific prosodic unit, and conveys a specific informative intention. The model provides a list of all possible factors that affect the prosody, and consequently the interpretation, of every IU.

2.1. The CALLIOPE space

CALLIOPE defines a multidimensional "space", where each "dimension" represents a characteristic influencing the vocal paralinguistic components of IUs; each characteristic is actually a categorical variable, assuming values in a set of labels. Each IU is thus associated to a "point" into this space; more formally, a generic IU is associated to a tuple $T_{(IU)}$ composed of 12 labels:

$$T_{(IU)} = (l_1, l_2, ..., l_{12}) : l_i \in F_i, 1 \leq i \leq 11. \quad (1)$$

Not all combinations are possible or commonly used: for example, it's not possible to utter a sentence like "*Giovanni ama Maria!*" ("Giovanni loves Maria!") as exclamatory and ironic at the same time [11]; this means that several tuples are never observed, or (which is the same) our space contains several points where no IUs can be associated. On the other hand we argue that, for a generic IU, it is possible to select a precise label for each dimension of our model; this means that every IU is truly a point in our space.

2.2. The CALLIOPE dimensions

CALLIOPE dimensions are divided into two groups. The first one contains Dialogic dimensions: characteristics directly related to the communication context; the corresponding F sets are fully defined. The second group contains Background dimensions: characteristics existing regardless of the presence of interaction; the corresponding F sets are finite, but so large that we prefer to consider them as open sets.

CALLIOPE defines the following Dialogic dimensions:

- F_1, Structure: *Declarative*, *Interrogative with 1 tonal unit*, *Interrogative with 2 or more tonal units*, *Interroga-*

10.21437/SpeechProsody.2018-65

tive disjunctive, *Echo questions*, *Exclamative*, *Vocative*. The labels were defined according to [12]. Note that the IU structure does not always correspond to the same intonation for all languages. For example, not all languages use final intonation rising to indicate a question.

- F_2, Linguistic modality: 24 labels divided into five main groups, as explained in [13]. According to [14]: "Modality is the category of meaning used to talk about possibilities and necessities, essentially, states of affairs beyond the actual". Examples of labels are *Advice*, *Suggestion*, *Opinion*.
- F_3, Intonational focus: *Presentational*, *Contrastive*, *Counter-presuppositional*, etc. These labels have been extracted from [15] and carry a pragmatic function. Such labels mainly fit intonational languages; about tonal languages, we are aware that there are ongoing studies about the influence of intonational focus [16].
- F_4, Rhetorical form: we consider only rhetorical forms that, we argue, influence the prosody of the sentence: *Irony*, *Aposiopesis*, *Prepetition*. We also included *Non-rhetorical*, for IUs lacking any rhetorical form.
- F_5, Motivational state: in the SMI (Interpersonal Motivational Systems) theory [17], motivations within interpersonal exchanges are analysed in an evolutionary perspective, taking into account psychotherapeutic dialogues. We can attribute to each IU a motivational label as shown in the AIMIT [18] manual for *Dominance/Submission*, *Care-seeking*, *Sexual bonding*, and *Caregiving* motivational systems.
- F_6, Speech mood: *Whispered*, *Normal*, and *Shouted speech*. We argue that such labels permit to represent common situations. Notice that the mood is often confused with Emotions and Linguistic modality. In our opinion, however, the mood represents a distinct, independent characterisation of the situation where the dialogue is situated [19, 20].
- F_7, Spontaneity: *Spoken*, *Read*, and *Recited* speech. These labels reflect the typologies found in speech corpora.
- F_8, Punctuation forms: they are a clue of the presence of pauses inside a IU. Notice that sometimes punctuation is used to separate different IUs, but in our case we are considering only punctuation inside a single IU. Examples are the labels *list* and *bracketing commas*.
- F_9, Emotion: as in [21], we choose: *Anger, Sadness, Fear, Joy, Love, Surprise, Disgust* and *Shame*. We are aware that classifying emotions is controversial and no definite list exists [22]. We argue, however, that the group we propose –together with Linguistic modality– could describe the vast majority of cases.

CALLIOPE defines the following Background dimensions:

- F_{10}, Subjective expressiveness skill: it's the level of personal skills the person shows during the dialog (e.g. *Able-bodied speaker*). It does not indicate the level of schooling, which –we argue– does not affect the prosody but only the linguistic skills.
- F_{11}, Social context: some special social contexts affect speakers' prosody. For example, a nursery teacher speaking with children, a politician explaining his plan in public, or a priest reciting a religious litany have different, emphasised or flattened prosody.
- F_{12}, Language, dialect or local accent form (e.g. *Standard Italian*): According to the main Universals in Language [23], speech derives from the need to express global concepts (e.g., talking about past or future events) that are common to all humans. Such concepts, however, have different effects on different linguistic systems [24]. For example, the sound of sarcasm changes among different languages [25]. Thus, we need to add to our model an explicit dimension where we represent each language variety.

3. The SI-CALLIOPE corpus

SI-CALLIOPE is a sub-space of CALLIOPE, defined by the following values for the Background dimensions:

- F_{12}=Standard Italian.
- F_{10}=Able-bodied speaker.
- F_{11}=Daily situation.

and the following values for the Dialogic dimensions:

- F_4=Non-rhetorical.
- F_6=Normal mood.
- F_7=Recited speech.

SI-CALLIOPE was defined according to the requirements of the LYV project. The purpose of the LYV project is to create a set of vocal interactive stories for persons with autism and other cognitive disabilities, based on vocal interaction and prosody skills. The interactive system will be able to automatically detect the prosody of the person and compare it with the general model.

We are aware that speech in autism can present problems like monotonic, creepy, whispering voice and a general low level degree of self-expressiveness linked to more specific psychological problems. Thus, we decided to start from a narrow group of simple, useful prosodic forms, which could be extended in future works.

3.1. The prosodic corpus: a subspace of SI-CALLIOPE

We recorded a set of sentences in collaboration with professional readers and actors. It was not possible to record sentences for all the SI-CALLIOPE space, as the number of possible combinations was still too big. Thus, we decided to focus on a subspace of SI-CALLIOPE, containing a narrow group of frequently-used prosodic forms, useful in the clinical, multimodal training sessions of the LYV project.

3.2. Recording sentences: the script

The script to read was presented as a list of sentences; each sentence was paired with the corresponding pseudo-sentence (see Section 3.4). Sentences were chosen to reflect common daily situations. Whenever a sentence interpretation was not clear, a short description of the context was added to the script.

In particular, the script is composed of 278 sentences (considering both real and pseudo-sentences) divided into the following labels.

- F_1, Structures: Declarative, Interrogative 1 tonal unit, Interrogative 2 or more tonal units, Interrogative disjunctive, Echo questions, Exclamative, General vocative.
- F_3, Intonational Focus: Contrastive Focus.

- F_4, Rhetorical form: Suspended.
- F_6, Speech mood: Whispered, Screamed, Normal.
- F_8, Punctuation forms: Lists.

Each label corresponds to a group of audio. To avoid too long recording session, we didn't include all possible combinations, focusing on the most common and significant ones.

3.3. Speaker characteristics and recording setting

We chose speakers without local or dialectal accents that could influence the global sound of the sentence, so we limited the group to people with experience in diction. We also chose recited speech because we needed clear, well-pronounced prosody samples.

We collected samples from 14 speakers (7 males and 7 females) 33 to 48 years old. All of them were experienced readers and actors.

Five speakers recorded sentences using their own equipment, while the remaining ones were recorded by one of the authors. Notice that it is actually preferable to collect recordings made with different equipment, in order to extrapolate a model that is as general as possible. Each speaker recorded 278 sentences (139 real-sentences and 139 pseudo-sentences) for a grand total of 3892 sentences.

The audio files were saved as WAV files (32-bit float, sample rate 44.1 kHz) using the software Audacity.

3.4. Generating pseudo-sentences

Our experiment required to create pseudo-sentences: sentences composed of words not belonging to the Italian dictionary, but generated so that they "sound" like Italian words.

To generate pseudo-sentences, we started from the CoLFIS corpus of Italian words [26], where we removed every word containing characters in the $\{w, y, j, k, x\}$ set, and every word containing characters with diacritical signs different from acute and grave accents. Then, the remaining words where split into syllables by means of Hyphenator 0.5.1 [27], a Python module that leverages the OpenOffice hyphenation dictionary. Finally, we trained a trigram of syllables that thus encoded an approximation of the Italian phonotactic.

Given a real-sentence, the algorithm generated, for each word longer than three characters, a random pseudo-word composed of the same number of syllables. The random generation leveraged the trigram so that the resulting pseudo-word respected Italian phonotactic. Words shorter that four characters (mostly articles, prepositions, and some instances of the verb "to be") were leaved unchanged to improve the readability of the sentence. The resulting pseudo-sentences were further cleaned to improve readability; in particular, a set of rules fixed genre and number concordances between articles/prepositions and the immediately following words.

Finally, pseudo-sentences were manually checked to further improve readability, and tonic accents were explicitly added. As an example, from the sentence "*Domani è bel tempo!*" we obtained the pseudo-sentence "*Selèzio è bel àmmi!*".

4. Psychoacoustic experiment

We are currently using SI-CALLIOPE in a perceptual experiment involving adult Italian native speaker (children are excluded because the construction of the prosodic model could be not fully developed.)

The test is available at `http://calliope.deib.polimi.it`. So far 253 tests have been collected, but the acquisition is still going on.

The main goal of the experiment is to understand the clues our brain needs to recognise a prosody; in particular, we aim to distinguish between:

- Prosodies that are recognised only thanks to basic acoustic clues (pitch contour and intensity being the most prominent ones). For these prosodies, the pitch contour could suffice for recognition. For example, we expect that the prosody of the Interrogative structure belongs to this category because of the typical rising of the pitch at the end of the sentence.
- Prosodies that are recognised thanks to acoustic and phonotactic. These prosodies need cues coming from the "sound" of syllables, even without any actual meaning. Complex acoustic features (i.e., based on spectrum, noise ratio, etc.) will be required to recognise such prosodies.
- Prosodies that need acoustic and actual meaning. These prosodies are actually a combination of sound cues and interpretation cues. Acoustic- and text-based features will be required to recognise such prosodies.

The experiment consists of 36 tests, divided into three sections:

- Section 1: 13 tests based on real sentences.
- Section 2: 13 tests based on pseudo-sentences.
- Section 3: 10 tests based on pitch-envelopes only. This section lacks the three tests concerning the perception of Speech mood, because it is impossible to detect it using only pitch envelopes.

For the generation of pitch envelopes we used the Praat [28] command `to Pitch`, smoothing with a value of 5Hz, and finally generating a sound with the `Hum` function. Such a function generates a sound similar to '*a*'. A slight fade-in was then applied to the obtained sounds to avoid the presence of artefacts that could disturb the listening.

In each test, the listener hears three IUs, and check which ones meet the test question; for example for the test "Which of the following audios is a direct question?", the listener will hear some questions and some non-question IUs, and should only check the formers.

Audios are extracted from the corpus with randomisation techniques, so that each person faces different IU combinations. In particular, for each test, the system inserts m_{yes} audio that the listener should check and $m_{no} = 3 - m_{yes}$ audio that the listener should discard, where $m_{yes} \in \{1, 2, 3\}$ is randomly chosen.

We are also collecting the following personal information: age, gender, and the subject's accent by selecting region and province. Accents of Italian speakers vary a lot, even inside the same province, so we are aware that our classification is only an approximation. Such information will be used for investigating whether age, gender, and accent affect the accuracy of listeners in recognising the prosodies.

5. The classifier

We are developing a Neural Network-based (NN) recogniser (see Figure 1) with the TensorFlow framework. The recogniser is meant to classify the three simplest prosodies: *Declarative*, *Interrogative (1 tonal unit)* and *Exclamative*.

5.1. Structure

From the psychoacoustic experiment, we understood that recognising prosodies needs a combination of meaning and acoustic clues. Thus, we designed our recogniser with three different NNs. The Audio-based NN leverages audio features (currently, we are using: pitch contour, intensity, their first and second derivatives, and some cepstrum bands). The Text-based NN is provided with the text –generated by an Automatic Speech Recognition (ASR) tool– the speaker is supposed to pronounce. Finally, the Master NN combines the predictions of both NNs and a softmax layer selects the most probable class.

In use cases where the voice quality is expected to be too low for the ASR to generate good transcriptions, the Audio-based NN can be directly connected to the softmax layer.

In uses cases where the goal is to evaluate speakers' skills in uttering correct prosodies (a typical scenario of the LYV project, where the expected prosody is known a priori), the Text-based NN is substituted with the expected prosody.

Both the Audio-based NN and the Text-based NN are based on a multi-layer, convolutional, Bidirectional Recurrent NN (C-BRNN), where each network node is a Long Short-Term Memory (LSTM) block.

5.2. Training the classifier: the ExInDe corpus

The SI-CALLIOPE corpus was too small for training a complex classifier. Thus, we built a new corpus, focusing on three simple prosodies that are easy to discriminate in a text (by means of the final punctuation mark): Exclamative, Interrogative with 1 tonal unit, and Declarative.

Starting from eBooks, ePub audio-eBooks, and various textual corpora, we extracted sentences, classified them by using the final punctuation mark, removed any punctuation mark[1] and obtained about 1.5 million of *textual IU samples*.

Starting from audio-eBooks (containing sentence-aligned audio and text), and various sentence-aligned textual/audio corpora we extracted about 60 000 *acoustic IU samples* of recited and spontaneous speech (and also added the corresponding text to the collection of textual IU samples.) Each acoustic/textual pair was classified using the same procedure explained above.

Notice that we designed ExInDe as a balanced corpus: each prosody is represented by about the same number of samples.

The Text-based and the Audio-based NNs were trained independently (attaching each of them to a softmax layer, to produce the classification) using, respectively, the textual IU samples and the audio part of the acoustic IU samples.

The Master NN, which is still under development, will be trained with audio and text of the acoustic IU samples; we will give to the NN both coherent samples (text and audio belonging to the same prosody) and contradiction samples (text and audio belonging to different prosodies); this way, the Master NN will learn how to evaluate the output of the two lower NNs.

6. Results

For the psychoacoustic experiment, we got data from 253 participants. Figure 2 shows the accuracy in recognising an Interrogative phrase. Surprisingly, it seems that hiding the meaning heavily affects the listeners' ability to recognise interrogative IUs as such; in fact, the accuracy of pseudo-sentences is more than 20% lower than the one of real-sentences. Moreover, the pitch envelope seems to confuse the listener (an accuracy below 50% is worst than a random choice); we argue that the signal produced by the Praat `Hum` command sounded a bit too unnatural and was hard to understand. To verify that the differences we found were real, we compared each of the three accuracies against the other two, and calculated the t-test. We obtained confidence values $1 - p > 0.999$ for all the comparisons. We are currently investigating the results we obtained, for the whole set of 13 tests composing the experiment.

For what concerns the automatic recognition we obtained an accuracy of about 67% for the Audio-based NN and 80% for the Text-based NN (as mentioned above,the Master NN is still under development.) Those results, however, are preliminary as we are still investigating about the best set of features to use.

[1] Keep in mind that the ExInDe texts should simulate ASR-generated transcriptions, so punctuation marks are not included.

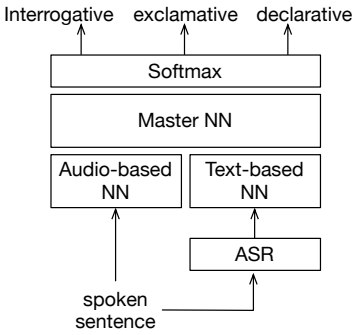

Figure 1: *The NN-based prosody classifier.*

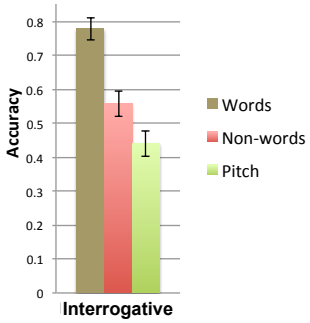

Figure 2: *Interrogative (error bars: binomials at 95%).*

7. Conclusions and future works

This paper described CALLIOPE, a new model for the categorisation of IUs; SI-CALLIOPE, a corpus of recorder prosodic forms, which is leveraged by an on-line perceptual experiment; and a NN-based model and corpus for prosody recognition.

We conducted a perceptive experiments with Italian speakers, discovering that recognising prosody, even for the simple Interrogation form, requires a combination of meaning and acoustic clues. This result implies that automatic prosody classification needs acoustic features as well as textual features.

We are still collecting data from the experiment, to confirm our findings and provide a complete statistical analysis; in particular, we will investigate whether gender, accent, and age affect prosody recognition. Moreover, we are still working on and improving the NN-based recogniser. Finally, we plan to extend the ExInDe corpus, to be able to train the classifier for recognition of more complex prosodic forms.

8. References

[1] J. Cole, "Prosody in context: a review," *Language, Cognition and Neuroscience*, vol. 30, no. 1-2, pp. 1–31, 2015.

[2] R. L. Mitchell, A. Jazdzyk, M. Stets, and S. A. Kotz, "Recruitment of language-, emotion-and speech-timing associated brain regions for expressing emotional prosody: investigation of functional neuroanatomy with fmri," *Frontiers in human neuroscience*, vol. 10, p. 518, 2016.

[3] S. Cenceschi, L. Sbattella, and R. Tedesco, "Verso il riconoscimento della prosodia," in *Studi AISV 2*, vol. 2, 2018.

[4] L. Sbattella and S. Guinea. (2016). [Online]. Available: http://www.polisocial.polimi.it/

[5] L. Sbattella, *La mente orchestra*. Vita e Pensiero, 2006.

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

[7] A. Beke and G. Szaszák, "Combining nlp techniques and acoustic analysis for semantic focus detection in speech," in *Cognitive Infocommunications (CogInfoCom), 2014 5th IEEE Conference on*. IEEE, 2014, pp. 493–497.

[8] S. Kori, E. Farnetani, and P. Cosi, "A perspective on relevance and application of prosodic information to automatic speech recognition in italian," in *European Conference on Speech Technology*, 1987.

[9] F. Tamburini, "Automatic prosodic prominence detection in speech using acoustic features: an unsupervised system," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[10] E. Cresti, *Corpus di italiano parlato: Introduzione*. Accademia della Crusca, 2000, vol. 1.

[11] M. D'Imperio, "Italian intonation: An overview and some questions," *Probus*, vol. 14, no. 1, pp. 37–69, 2002.

[12] P. Prieto, J. Borràs-Comes, and P. Roseano, "Interactive atlas of romance intonation," *Web page: http://prosodia. upf. edu/iari*, 2010.

[13] A. Kratzer, *Modals and conditionals: New and revised perspectives*. Oxford University Press, 2012, vol. 36.

[14] C. Maienborn, K. von Heusinger, and P. Portner, *Semantics: An international handbook of natural language meaning*. Walter de Gruyter, 2011, vol. 1.

[15] C. Gussenhoven, "Types of focus in english," in *Topic and focus*. Springer, 2008, pp. 83–100.

[16] I. C. Ouyang and E. Kaiser, "Focus-marking in a tone language: Prosodic cues in mandarin chinese," in *LSA Annual Meeting Extended Abstracts*, vol. 3, 2012, pp. 8–1.

[17] G. Liotti and F. Monticelli, "I sistemi motivazionali nel dialogo clinico," *Il manuale AIMIT*, 2008.

[18] G. Fassone, F. Valcella, S. Pallini, F. Scarcella, L. Tombolini, A. Ivaldi, E. Prunetti, F. Manaresi, and G. Liotti, "Assessment of interpersonal motivation in transcripts (aimit): an inter-and intrarater reliability study of a new method of detection of interpersonal motivational systems in psychotherapy," *Clinical psychology & psychotherapy*, vol. 19, no. 3, pp. 224–234, 2012.

[19] P. N. Juslin and K. R. Scherer, "Speech emotion analysis," *Scholarpedia*, vol. 3, no. 10, p. 4240, 2008.

[20] K. R. Scherer, "What are emotions? and how can they be measured?" *Social science information*, vol. 44, no. 4, pp. 695–729, 2005.

[21] S. S. Tomkins, "Affect theory," *Approaches to emotion*, vol. 163, no. 163–195, 1984.

[22] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech communication*, vol. 40, no. 1, pp. 5–32, 2003.

[23] S. Cristofaro, *Language universals and linguistic knowledge*. na, 2006.

[24] J. M. Sadock and A. M. Zwicky, "Speech act distinctions in syntax," *Language typology and syntactic description*, vol. 1, pp. 155–196, 1985.

[25] H. S. Cheang and M. D. Pell, "Acoustic markers of sarcasm in cantonese and english," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1394–1405, 2009.

[26] P. M. Bertinetto, C. Burani, A. Laudanna, L. Marconi, D. Ratti, C. Rolando, and A. M. Thornton, "Corpus e lessico di frequenza dell'italiano scritto (colfis)," *Scuola Normale Superiore di Pisa*, 2005.

[27] W. Berendsen. (2013). [Online]. Available: https://pypi.python.org/pypi/hyphenator/0.5.1/

[28] P. Boersma and D. Weenik, "Praat: a system for doing phonetics by computer. report of the institute of phonetic sciences of the university of amsterdam," *Amsterdam: University of Amsterdam*, 1996.