

CALLIOPE: a multi-dimensional model for the prosodic characterisation of Information Units

Sonia Cenceschi¹[0000-0002-4145-9593], Licia Sbattella²[0000-0001-5344-5976], and
Roberto Tedesco²[0000-0002-2830-4247]

¹ Digital forensic service, Department of innovative technologies, University of
Applied Sciences and Arts of Southern Switzerland (SUPSI), Manno, Switzerland

`sonia.cenceschi@supsi.ch`

`http://www.supsi.ch/forensic_en.html`

² ARCSLab, Department of Electronics, Information and Bioengineering
Politecnico di Milano, Milan, Italy

`{licia.sbattella,roberto.tedesco}@polimi.it`

Abstract. CALLIOPE is a conceptual multi-dimensional model that aims at approximating and categorising the prosodic phenomena. Its goal is to provide a method suitable for the analysis of dialogue and dialogic interaction, and the design of vocal corpora usable in the development of deep learning algorithms, merging theories belonging to different research fields on human language. CALLIOPE contains all possible independent factors affecting the sound of so-called Information Units (IUs). In CALLIOPE, each IU is associated with a tuple composed of 12 labels, where each of them belongs to a different dimension representing a characteristic influencing the prosodic behaviour. These dimensions are divided into two groups: the first one containing fully defined characteristics directly related to the dialogic dimension (e.g. Motivational state, Rhetorical form, etc.), while the second one takes into account dimensions that exist regardless of the presence of interaction (e.g. Language or dialect, social context, etc.).

Keywords: Prosody · Prosodic model · Information Unit

1 Introduction

CALLIOPE (Combined and Assessed List of Latent Influences On Prosodic Expressivity) starts from the need to characterise the acoustic variability of human speech, with the ultimate aim of creating corpora suitable for the analysis of dialogue and dialogic interaction, and the development of deep learning algorithms. In fact, creating an appropriate and well-defined prosodic corpus is a substantial condition for the correct functioning of automatic algorithms, which otherwise risk to provide unbalanced results. CALLIOPE focuses on all the factors co-constructing a shared dialogical act, and in particular on the relationship between linguistic, phonotactical and intonational parts of so-called Information

Units (IU). Following [14], a given IU is composed by a *textual realisation* (i.e., a written phrase) and an *acoustic realisation* (an audio recording of a speaker uttering such a phrase), and conveys an *Illocutionary Act* [1]. The acoustic realisation carries both phonotactics and prosody. Phonotactics is the possible arrangement of phones in words of a given language while, according to [25], prosody is the systematic organisation of various linguistic units into an utterance in the process of speech production, whose realisation involves both segmental and suprasegmental features.

Humans establish a dialogue relationship with others through many different paralinguistic clues: kinesthetics, environmental context, and other exogenous complex linguistics and spoken signals. In this perspective, the acoustic component (in other words, the prosody) plays a crucial role. Consider, for example, *irony* and *focus*: with the first, the messages conveyed by the sentence and its acoustic realisation are in contradiction creating a refined pragmatic game, while the second can be used to introduce a new element in the discussion, or to underline the conclusion of an intervention within the conversation. Moreover, the understanding of a spoken message does not always coincide with the information conveyed by the linguistic content [12] and, at the same time, it can be decomposed into many aspects not limited to single word processing.

1.1 State of the art

To our knowledge, no model allowing IUs' prosody categorisation has been proposed, and moreover, models taking into account the many variables involved in the communication process. This work contains a first proposal obtained by merging different theories, belonging to the most varied fields of study. Looking at the pertaining works, many of them are insightful but often represent partial, different or incomplete points of view. The most relevant theory is the Language into Act theory (L-AcT) described in [15], a Speech Act theory [1] extension focused on the pragmatic sense of the production of IUs. CALLIOPE is based on the Illocutionary Act classification provided by this theory.

CALLIOPE has also been inspired by The Interactive Atlas of Romance Intonation (IARI) [46] focused on the study of intonation in different Romance languages. In particular, we relied on the script IARI defines for sample recordings. IARI uses a survey for eliciting semi-spontaneous speech, providing useful but prosodically incomplete clues. IARI leverages the ToBi system [3] for tagging intonation (ToBi is a set of conventions for transcribing and annotating some prosodic phenomena, intonation being the most prominent one.) However, ToBi and other intonational transcription systems are based on manual annotations, which can be very unreliable if listeners are not experts [31, 53]; so, we didn't make use of it. Instead, the work done by Wennerstrom in [52] is more interesting and complete, as it integrates the ToBi notation in dialogue interaction analysis (e.g. the process of concordance and discordance of the interlocutors).

We adopted the general intonation structure defined by IARI: Statements, Yes-no questions, Wh-questions, Echo questions, Imperatives and Vocatives. A statement is commonly considered a declarative sentence, but we considered just

its intonation and assigned the purpose to its contextualised Illocutionary Act. Consider the two IUs *Domani è bel tempo* (“Tomorrow the weather will be fine”) and *DOMANI è bel tempo* (“TOMORROW the weather will be fine”); while the first IU is informative, the second one is intended to correct a previous speaker.

Prosody is also a mean for communicating emotions, and so we need to take into account that the acoustic realisation of the IU could carry an affective valence [55]. Emotion categorisations are innumerable, but we relied on the well-known Plutchik and Tomkins theories [52, 45]. Emotions-speech relationship has been deeply investigated to reproduce the neurological detecting processes, and links between acoustic features and emotional states are widely studied; some examples can be found in [54, 43, 51].

Emotions are felt together with the action of the speakers into the dialogue according to particular goals and mutual interactions. This topic has been largely studied into the Interpersonal Motivational Systems (SMI) theory [38] from a psychological point of view. Detection of SMIs has been initially based on the analysis of textual contents (see the AIMIT Manual: [23]); nevertheless, sound plays a very important role. For example, the same request is pronounced differently if we want to induce the interlocutor to trust us, or if we want to assert some kind of social superiority. The ARCSLab³ has combined textual and acoustic components of the speech to automatically extract SMIs [48], based on a original multilevel model of verbal interaction, in order to analyze the evolution of forensic interrogations.

Another useful point of view in the classification of IUs is the placement of pragmatic stresses. The theory distinguishes between different types of stresses but here we restrict our analysis to the classification of *focus*, which is a particular emphasis placed in order to attract the listener’s attention to a specific part of the IU. In our classification we considered categories proposed by [30] but many other theories exist [7, 20, 26]. The way we intend the focus comes from the L-AcT theory, which underlines that the sound is a direct consequence of a pragmatic act.

Speech styles and context change prosody as well: read, recited and spontaneous speech lead to different prosodies [42, 39] and also position in space and context [32, 27].

Finally, phonotactics and acoustic phonetics should be considered. Each language defines restrictions on the feasible combination of phonemes leading to different sound and segmental or suprasegmental combinations. Phonotactics, and therefore the language variation, deeply influences prosody [6, 4]: this is why we will consider the idiom as a space where other parameters are varying.

Only works considering a direct connection with the prosodic realisation have been used in CALLIOPE. However, widening the horizon to theories regarding phenomena indirectly influencing IU acoustic realisations, we find, for example, the linguistic and expressive abilities, their typology and nature and the social

³ The ARCSLab laboratory (Adaptable, Relational and Cognitive Software Environments) directed by prof. L. Sbattella is based at the Dipartimento di Elettronica, Informazione e Bioingegneria at Politecnico di Milano. <https://arcslab.dei.polimi.it/>

context [?]. Many of these phenomena are essentially not measurable factors, and thus we decided not to consider them, but nevertheless they do affect IU acoustic realisations.

The influence of the linguistic meaning on human understanding processes has been largely explored [19, 34, 2] but focusing in particular on the neurological functioning. Focusing on prosody, [36] describe algorithms and statistical models leveraging prosodic information on various levels of speech understanding, while [44] investigate about how prosody can be used in automatic speech understanding systems.

2 The CALLIOPE model

CALLIOPE is a model aiming at categorising all prosodic forms. Each IU has a bi-univocal correspondence with a specific prosodic realisation, 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.

CALLIOPE defines a multidimensional “space”, where each *dimension* represents a characteristic influencing the acoustic paralinguistic components of IUs; each dimension is actually a categorical variable, assuming values in a set of labels called *elements*. 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 elements:

$$T(IU) = (l_1, l_2, \dots, l_{12}) : l_i \in D_i, 1 \leq i \leq 12. \quad (1)$$

Some combinations are not possible or commonly used: for example, it’s not possible to utter a IU like “*Giovanni ama Maria!*” (“Giovanni loves Maria!”) as exclamatory (D_1) and ironic (D_4) at the same time [22]; 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 element for each dimension of our model; this means that every IU is truly a point in our “space”.

CALLIOPE dimensions are divided into two groups. The first one contains *Dialogic dimensions*: characteristics directly related to the communication context; the corresponding D sets are fully defined. The second group contains *Background dimensions*: characteristics existing regardless of the presence of interaction; the corresponding D sets are finite, but so large that we prefer to consider them as open sets. Table 1 lists the dimensions for each group and shows the research field each dimension derives from. Each dimension is describe in detail in the following sections, while the complete list of CALLIOPE labels is provided in Appendix A.

2.1 Dialogic dimensions

This section details the CALLIOPE dialogic dimensions; all the following drawings have been generated with Praat [5].

Table 1. The CALLIOPE dimensions and the research fields they derive from

Dimensions	Field of study
Dialogic dimensions	
D ₁ STRUCTURE	Grammar & phonotactics
D ₂ ILLOCUTIONARY ACT	Pragmatics
D ₃ INTONATIONAL FOCUS	Pragmatics
D ₄ RHETORICAL FORM	Pragmatics
D ₅ MOTIVATIONAL STATE	Psychology
D ₆ SPEECH LOUDNESS	Spatial dislocation
D ₇ SPONTANEITY	Sociology
D ₈ ACOUSTIC PAUSE	Unforeseeable factors
D ₉ EMOTION	Psychology
Background dimensions	
D ₁₀ SUBJECTIVE EXPRESSIVENESS SKILLS	Clinics
D ₁₁ SOCIAL CONTEXT	Sociology
D ₁₂ LANGUAGE, DIALECT OR LOCAL ACCENT FORM	Grammar & phonotactics

D₁ (Structure) This dimension refers to the sentence typology, as determined by both a specific punctuation mark and a some peculiar prosodic characteristics. We adopted the following typologies: **Declarative**, **Interrogative with 1 tonal unit**, **Interrogative with 2 or more tonal units**, **Interrogative disjunctive**, **Echo question**, **Exclamative**, and **Vocative**. Such elements have been inspired by The Interactive Atlas of Romance Intonation [46].

Intonation (pitch envelope in time) is one of the main components of prosody and is often used to easily describe different structures (see Figure 1), but this is not sufficient to characterise and discriminate this dimension. The *Pitch* and *Intensity* envelopes, together, can provide a first approximated representation of the IU STRUCTURE; see Figure 2, where the two contours suggest the presence of a repetition. However, other spectral features contribute to prosodic variations and perception. Furthermore, note that a structure typology does not correspond to the same prosodic realisation in all languages; for example, not all languages use final intonation rising to indicate a question. This characteristic (which is actually present in several dimensions) forced us to add to CALLIOPE a specific dimension defining the current language: D₁₃ (LANGUAGE, DIALECT OR LOCAL ACCENT FORM).

D₂ (Illocutionary Act) This dimension derives from the concept of Speech Act [1], where “to say something is to do something” and an utterance is considered as an action. In this perspective, the issuing of an utterance in a speech situation changes with the intention of the speaker, corresponding to the his/her type of attitude, called ILLOCUTIONARY ACT (Examples are **Advice**, **Suggestion**, and **Opinion**). Another approach [41] defines it as “the category of meaning used to talk about possibilities and necessities, essentially, states of affairs beyond the actual”. In this work we rely on 26 ILLOCUTIONARY ACTS, following the categorisation explained in [37]. Notice that this is just one of many differ-

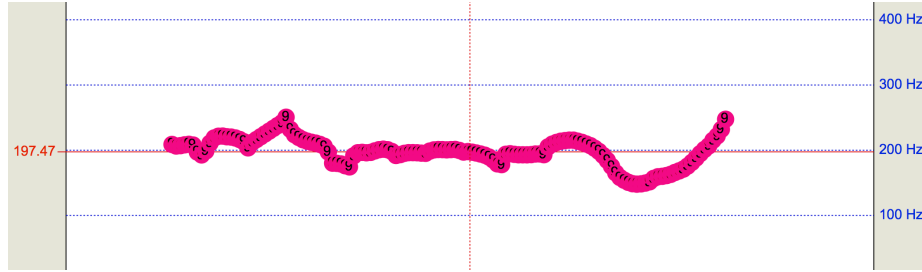


Fig. 1. Standard Italian intonation of a Question with 1 tonal unit, obtained with the Praat `Smooth` command.

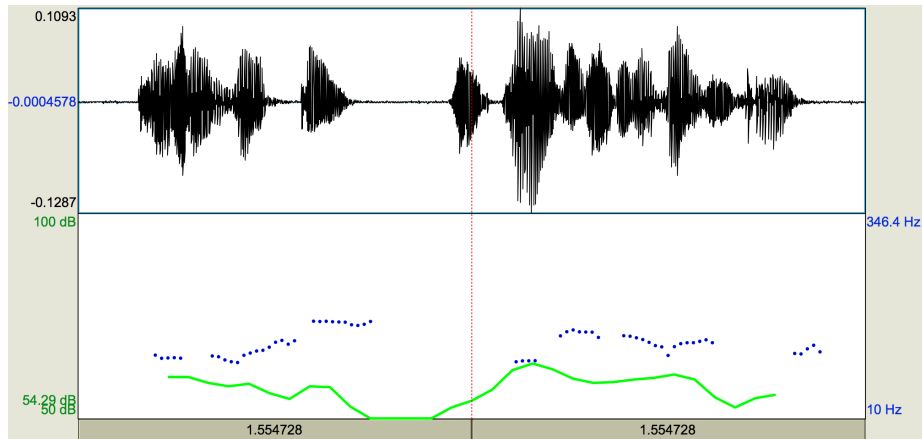


Fig. 2. Standard Italian Echo question: intensity (green) and pitch (blue dots) envelopes.

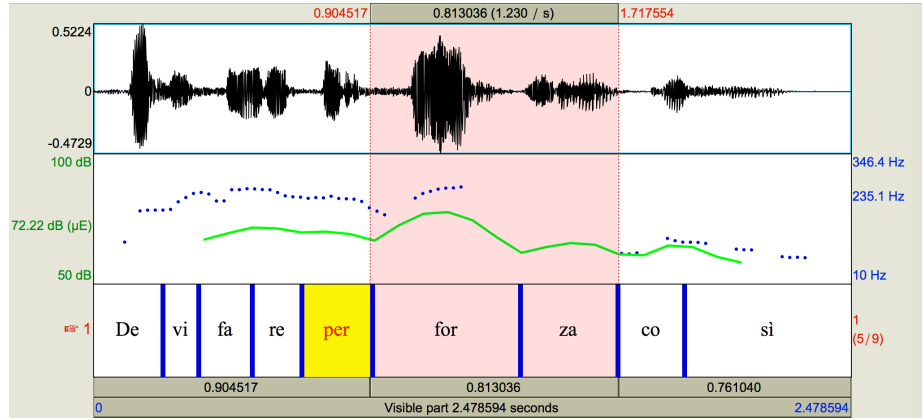


Fig. 3. Standard Italian **Corrective** focus. The IU is: *Devi fare per forza così!* (“You must do that!”) with the focus on the syllable *for-*.

ent categorisations, which can define as many as 90 acts [17, 24], extending the Speech Act theory to the theory of Language into Act (L-Act) for spontaneous speech [16].

D₃ (Intonational Focus) This dimension highlights important elements of the spoken message. The acoustic realisation of the focus depends on speaker’s culture and language [29]. For example, the Standard Italian language is strongly syllable-timed: syllables take approximately an equal amount of time to be pronounced, and they are temporally stretched by speakers when they intend to underline a word. A clear example is shown in Figure 3, where the focused word *forza* (in this context it means “you must...”) is the longest one and presents higher pitch and intensity. The focused syllable is characterised by changes in fundamental frequency excursion and intensity-related parameters, with respect to their average values. Listener’s expectations affects how prominence is perceived [49].

In our model we adopted elements like **Presentation**, **Corrective**, **Counter-suppositional**, etc. These elements have been extracted from [30] and they carry a clear pragmatic function. The full list, as well as the precise semantics, is shown in the Appendix A. We also included **Non-focused**, for IUs lacking any **INTONATIONAL FOCUS**.

D₄ (Rhetorical Form) We consider only the rhetorical forms [33, 21] that are detectable in acoustic realisation of IUs: **Irony**, **Aposiopesis**, **Prepetition**, **Anacoluton**, etc. We also included the **Null** element for IUs lacking any rhetorical form. The full list, as well as the precise semantics, is shown in Appendix A. Simple examples are the **Enumeratio** (e.g. “1, 2, 3, ...”), where commas insert temporal pauses; **Irony**, where the acoustic realisation can communicate an opposite effect

with respect to the linguistic part of the IU; and **Aposiopesis**, where a sentence is deliberately broken off and left unfinished (e.g. “Get out, or else...!”).

The **Dialisis** and **Parenthesis** elements are two particular cases: both of them are made of two different sentences joined together and considered as a single IU (*Ho scordato, ma chi se ne importa, l’ombrello* - “I forgot, but who cares, the umbrella”).

D₅ (Motivational State) In the SMI theory, motivations within interpersonal exchanges are analysed in an evolutionary perspective. The AIMIT manual defines the following SMIs: **Care-seeking**, **Sexual bonding**, **Caregiving**, **Attachment**, **Competitive**, **Peer-cooperative** and **Social rank**. We also included the **Neutral** element, for IUs where no SMI is detectable. It is important to underline that this dimension differs both from **SOCIAL CONTEXT** and **ILLOCUTIONARY ACT**; in fact, SMIs are related to the relationship between the interlocutors and their mutual role. In each different **SOCIAL CONTEXT** speakers can independently play different interpersonal roles (e.g. the judging and the judged) and inside this SMI they can choose one of the 26 **ILLOCUTIONARY ACT** elements (e.g. both the judging person and the judged person can use IUs to convince, suggest, ask, etc.) At the same time, the SMI modifies the acoustic realisation of the IU because prosody is influenced by the role we are playing with respect to our interlocutors (e.g. the judging person will be more secure and less undecided than the judged person, pronouncing the same IU). The SMI theory is admittedly very abstract; notice, however, that SMIs can be automatically extracted from IUs. As an example see the DIKE project [48], a pioneer work regarding automatic dialogue analysis exploiting an original multilevel model of verbal interaction taking into account of the mutual indivisible relationship between SMI and prosody.

D₆ (Speech Loudness) This dimension depends on the situation where the dialogue is happening. We named this dimension after the *loudness* classic definition: the subjective perception of sound pressure (intensity). In this case, it is used with a wider sense and it does not only affect intensity: a speaker could pronounce the same IU with low, normal or high intensity, trying to maintain the same global prosodic behaviour, but the result would sound very unnatural. In fact, **SPEECH LOUDNESS** affects several acoustic features, generating different prosodic realisations. For example, think about talking with someone positioned quite far apart or close to you: it is immediately clear that the same IU is expressed by very different acoustic realisations. Such differences are highlighted by Figure 4. We define the following elements: **Whispered**, **Normal**, and **Shouted speech**. It must be underlined that **Whispered speech** has specific acoustic properties, which interact in a particular way with the other dimensions and the extraction of acoustic features such as pitch or formants.

D₇ (Spontaneity) This dimension reflects the speech typologies found in corpora [42], which reflect three main types of enunciative styles: **Spoken**, **Read**, and

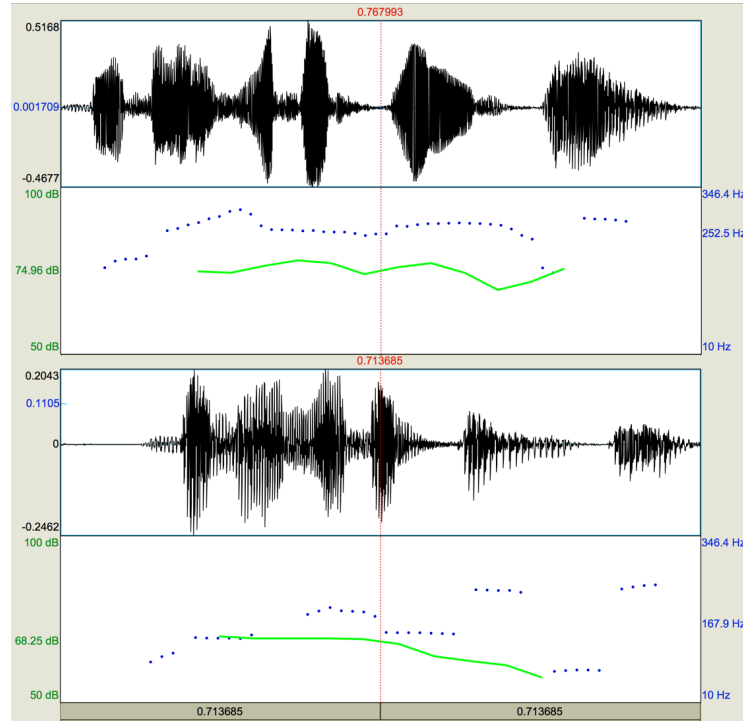


Fig. 4. The same speaker pronouncing the same IU, screaming (above) and with a normal loudness (below): *Domani è bel tempo!* (“Tomorrow the weather is fine!”)

Recited speech. Like in EMOTION, such element set is potentially expandable but we think that, together with ILLOCUTIONARY ACT and SOCIAL CONTEXT, we can provide a good approximation for all the IUs.

D₈ (Acoustic Pause) This dimension describes the presence of purely acoustic pauses inside an IU. This kind of pauses lack any pragmatic valence and is due to extemporaneous phenomena (or, like in poetry, for creating particular effects). Often, at the textual realisation level, such pause is represented by a comma or three dots (e.g., *Vieni, a bere un caffè* - “Come, to drink a coffee”). Elements are **Present** and **Not Present**. Notice that bracketing commas and the incidental sentence are not an ACOUSTIC PAUSE; instead, they are considered by the Parenthesis and Dialysis elements of RHETORIC DIMENSION.

D₉ (Emotion) We are aware that classifying emotions is controversial and no definite list exists [13]. As in [50], we choose to consider the following basic ones: Anger, Sadness, Fear, Joy, Love, Surprise, Disgust and Shame. The group we propose –together with ILLOCUTIONARY ACT– could describe the vast major-

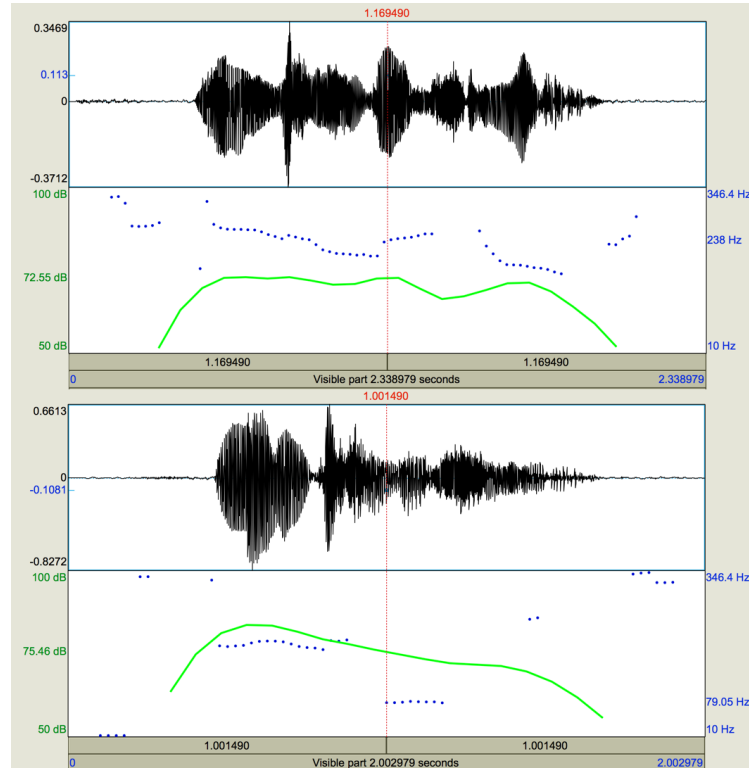


Fig. 5. The same speaker pronouncing the same IU with sadness (above) and anger (below): “*Non voglio uscire!*” (“I don’t want to go out!”)

ity of cases. We also included the **Neutral** element for IUs lacking a particular emotional state. Figure 5 shows an example.

2.2 Background dimensions

This section details the CALLIOPE background dimensions.

D₁₀ (Subjective Expressiveness Skills) This dimension describes the expressiveness skills the person shows during the dialogue. For example, a speaker with cognitive problems influencing the expressiveness level may not complete the words, or have a flat and monotonous intonation. It is not possible to define a comprehensive list of elements because there are endless variations and subjective nuances: that’s why D₁₀ is a Background dimensions with an open set of elements.

D₁₁ (Social Context) Some special social contexts affect speakers’ prosody [35]. For example, a nursery teacher speaking with children, a politician explaining his

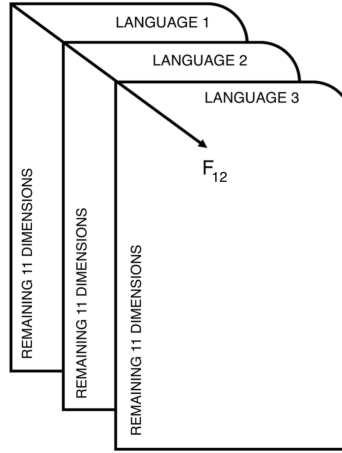


Fig. 6. How to visualise Language dialects or local accent dimensions.

plan in public, or a priest reciting a religious litany have different, emphasised or flattened prosody. Indicating the speaker’s **Social Context** helps in ensuring the repeatability of experiments, and permits to accurately contextualise the research (e.g. during an experimental research concerning the persuasive ability of the political speech). It is not possible to provide a comprehensive list of elements, and this is the reason why D_{11} is a Background dimensions with an open set of elements.

D_{12} (Language, Dialect or Local Accent Form) According to the main Universals in Language [18], 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 [47]. For example, the acoustic realisation of sarcasm changes among different languages [11]. Thus, we added to our model an explicit dimension where we represent each language variety. In this sense, each language variety can be imagined as a space “slice”, like the ones shown in Figure 6, where in each one of them the phonotactics [28] and the expressiveness modalities of the specific community are applied.

Indicating the speaker’s **LANGUAGE, DIALECT OR LOCAL ACCENT FORM** helps in ensuring the repeatability of experiments, and permits to accurately contextualise the research (e.g. in a researches concerning how local accents affect the acoustic realisation of the yes/no question).

We do not list all the languages and consider D_{12} as an open set of elements.

3 Conclusion and future work

CALLIOPE is a very general and conceptual model, and is too big to be validated as a whole. Then, a first attempt has been made working on SI-CALLIOPE, a sub-space based on recited Italian speech. Basing on it, we built a speech corpus used for psycho-acoustic experiments of partial validation, and automatic prosodic features recognition (see works [9, 10, 40, 8]). Future work will explore the independence of the listed dimensions and analyse the interaction between the various labels. The fundamental role of CALLIOPE is the characterization of each IU with a single tuple according to twelve dimensions, but nothing prevents to expand the labels set for one or more of them. For example, EMOTION, ILLOCUTIONARY ACT or SPEECH MOOD can be refined using more labels. Therefore, to meet the CALLIOPE's requirements, each new label should be analysed to define its influence on prosody, and validated with perceptual experiments to prove its independence from others.

References

1. Austin, J.L.: How to do things with words, vol. 88. Oxford university press (1975)
2. Baker, J.M., Deng, L., Glass, J., Khudanpur, S., Lee, C.H., Morgan, N., O'Shaughnessy, D.: Developments and directions in speech recognition and understanding, part 1 [dsp education]. *Ieee Signal processing magazine* **26**(3), 75–80 (2009)
3. Beckman, M.E., Hirschberg, J., Shattuck-Hufnagel, S.: The original tobi system and the. Prosodic typology: The phonology of intonation and phrasing **1**, 9 (2006)
4. Bennett, R.T.: Foot-conditioned phonotactics and prosodic constituency. Ph.D. thesis, UC Santa Cruz (2012)
5. Boersma, P.: Praat, a system for doing phonetics by computer. *Glott. Int.* **5**(9), 341–345 (2001)
6. Booij, G., et al.: The role of the prosodic word in phonotactic generalizations. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4* pp. 47–72 (1999)
7. Büring, D., et al.: Towards a typology of focus realization. *Information structure* pp. 177–205 (2009)
8. Cenceschi, S., Sbattella, L., Tedesco, R.: Influence of semantics on the perception of corrective focus in spoken italian. *ExLing* 2018 p. 29 (2018)
9. Cenceschi, S., Sbattella, L., Tedesco, R.: Towards automatic recognition of prosody. In: *Proc. 9th International Conference on Speech Prosody*. pp. 319–323 (2018)
10. Cenceschi, S., Sbattella, L., Tedesco, R.: Verso il riconoscimento automatico della prosodia. *STUDI AISV* pp. 433–440 (2018)
11. Cheang, H.S., Pell, M.D.: Acoustic markers of sarcasm in cantonese and english. *The Journal of the Acoustical Society of America* **126**(3), 1394–1405 (2009)
12. Cole, J.: Prosody in context: a review. *Language, Cognition and Neuroscience* **30**(1-2), 1–31 (2015)
13. Cowie, R., Cornelius, R.R.: Describing the emotional states that are expressed in speech. *Speech communication* **40**(1-2), 5–32 (2003)
14. Cresti, E.: *Corpus di italiano parlato: Introduzione*, vol. 1. Accademia della Crusca (2000)

15. Cresti, E.: Syntactic properties of spontaneous speech in the language into act theory. *Spoken Corpora and Linguistic Studies*. Amsterdam/Philadelphia: John Benjamins pp. 365–410 (2014)
16. Cresti, E.: Syntactic properties of spontaneous speech in the language into act theory. *Spoken Corpora and Linguistic Studies*. Amsterdam/Philadelphia: John Benjamins pp. 365–410 (2014)
17. Cresti, E., Martin, P., Moneglia, M.: L'intonazione delle illocuzioni naturali rappresentative: analisi e validazione percettiva. *Atti delle IX giornate del gruppo di fonetica sperimentale (AIA)*. Padova: Unipress pp. 51–63 (1998)
18. Cristofaro, S.: Language universals and linguistic knowledge. na (2010)
19. Davis, M.H., Johnsrude, I.S.: Hearing speech sounds: top-down influences on the interface between audition and speech perception. *Hearing research* **229**(1-2), 132–147 (2007)
20. Domínguez, L.: Mapping focus: The syntax and prosody of focus in Spanish. Ph.D. thesis, Boston University (2004)
21. Dupriez, B.M.: *A Dictionary of Literary Devices*: Gradus, AZ. University of Toronto Press (1991)
22. D'Imperio, M.: Italian intonation: An overview and some questions. *Probus* **14**(1), 37–69 (2002)
23. Fassone, G., Valcella, F., Pallini, S., Scarcella, F., Tombolini, L., Ivaldi, A., Prunetti, E., Manaresi, F., Liotti, G.: Assessment of interpersonal motivation in transcripts (aimit): an inter-and intra-rater reliability study of a new method of detection of interpersonal motivational systems in psychotherapy. *Clinical psychology & psychotherapy* **19**(3), 224–234 (2012)
24. Firenzuoli, V.: Le forme intonative di valore illocutivo dell'italiano parlato: analisi sperimentale di un corpus di parlato spontaneo (LABLITA): tesi di dottorato (2001)
25. Fujisaki, H.: Prosody, models, and spontaneous speech. In: *Computing prosody*, pp. 27–42. Springer (1997)
26. Fujisaki, H.: Information, prosody, and modeling-with emphasis on tonal features of speech. In: *Speech Prosody 2004, International Conference* (2004)
27. Ghaffarzadegan, S., Bořil, H., Hansen, J.H.: Model and feature based compensation for whispered speech recognition. In: *Fifteenth Annual Conference of the International Speech Communication Association* (2014)
28. Goldrick, M.: Phonological features and phonotactic constraints in speech production. *Journal of Memory and Language* **51**(4), 586–603 (2004)
29. Gundel, J., Bosch, P., van der Sandt, R.: *Focus: Linguistic, cognitive, and computational perspectives*. Cambridge: Cambridge University Press. doi **10** (1999)
30. Gussenhoven, C.: Types of focus in english. In: *Topic and focus*, pp. 83–100. Springer (2008)
31. Gut, U., Bayerl, P.S.: Measuring the reliability of manual annotations of speech corpora. In: *Speech Prosody 2004, International Conference* (2004)
32. Hansen, J.H., Nandwana, M.K., Shokouhi, N.: Analysis of human scream and its impact on text-independent speaker verification. *The Journal of the Acoustical Society of America* **141**(4), 2957–2967 (2017)
33. Harris, R., et al.: *A handbook of rhetorical devices*. Virtual Salt Salt Lake City, UT (2002)
34. Kazanina, N., Phillips, C., Idsardi, W.: The influence of meaning on the perception of speech sounds. *Proceedings of the National Academy of Sciences* **103**(30), 11381–11386 (2006)

35. Klasen, M., von Marschall, C., Isman, G., Zvyagintsev, M., Gur, R.C., Mathiak, K.: Prosody production networks are modulated by sensory cues and social context. *Social cognitive and affective neuroscience* **13**(4), 418–429 (2018)
36. Kompe, R., Kompe, R.: Prosody in speech understanding systems, vol. 1307. Springer (1997)
37. Kratzer, A.: Modals and conditionals: New and revised perspectives, vol. 36. Oxford University Press (2012)
38. Liotti, G., Monticelli, F.: a cura di (2008). I sistemi motivazionali nel dialogo clinico: il manuale AIMIT
39. Llisterri, J.: Speaking styles in speech research. In: Workshop on Integrating Speech and Natural Language (1992)
40. López Zorrilla, A., De Velasco Vázquez, M., Cenceschi, S., Torres Barañano, M.I.: Corrective focus detection in italian speech using neural networks (2018)
41. Maienborn, C.: Semantics: an international handbook of natural language meaning. 1 (2011), vol. 1. Walter de Gruyter (2011)
42. Nencioni, G.: Di scritto e di parlato, vol. 6. Zanichelli (1983)
43. Nicholson, J., Takahashi, K., Nakatsu, R.: Emotion recognition in speech using neural networks. *Neural computing & applications* **9**(4), 290–296 (2000)
44. Noth, E., Batliner, A., Kießling, A., Kompe, R., Niemann, H.: Verbmobil: The use of prosody in the linguistic components of a speech understanding system. *IEEE Transactions on Speech and Audio processing* **8**(5), 519–532 (2000)
45. Plutchik, R.: The emotions. University Press of America (1991)
46. Prieto, P., Borràs-Comes, J., Roseano, P.: Interactive atlas of romance intonation. Web page: <http://prosodia.upf.edu/iari> (2010)
47. Sadock, J.M., Zwicky, A.M.: Speech act distinctions in syntax. *Language typology and syntactic description* **1**, 155–196 (1985)
48. Sbattella, L., Tedesco, R., Trivilini, A.: Forensic examinations: Computational analysis and information extraction. In: International Conference on Forensic Science-Criminalistics Research (FSCR). pp. 1–10 (2014)
49. Tamburini, F., Bertini, C., Bertinetto, P.M.: Prosodic prominence detection in italian continuous speech using probabilistic graphical models. In: Proceedings of Speech Prosody. pp. 285–289 (2014)
50. Tomkins, S.S.: Affect theory. Approaches to emotion **163**(163-195) (1984)
51. Vogt, T.: Real-time automatic emotion recognition from speech (2010)
52. Wennerstrom, A.: The music of everyday speech: Prosody and discourse analysis. Oxford University Press (2001)
53. Wightman, C.W.: Tobi or not tobi? In: Speech Prosody 2002, International Conference (2002)
54. Williams, C.E., Stevens, K.N.: Emotions and speech: Some acoustical correlates. *The Journal of the Acoustical Society of America* **52**(4B), 1238–1250 (1972)
55. Zentner, M., Grandjean, D., Scherer, K.R.: Emotions evoked by the sound of music: characterization, classification, and measurement. *Emotion* **8**(4), 494 (2008)

A THE CALLIOPE LABELS

The full list of labels, for each CALLIOPE dimension.

F₁ - Structure

1. Declarative
2. Interrogative with 1 tonal unit
3. Interrogative with 2 or more tonal units
4. Interrogative disjunctive
5. Echo questions
6. Exclamative
7. Vocative

F₂ - Linguistic modality

25 labels divided into five main groups.

Alethic:

1. Assumption
2. Confirmation
3. Objection
4. Admission
5. Ascertainment
6. Description
7. Explanation
8. Clarification
9. Inference

Epistemic:

10. Intuition
11. Conjecture
12. Inference
13. Doubt
14. Supposition
15. Prediction
16. Query

Appreciative:

17. Opinion
18. Judgement

Volitive:

19. Desire
20. Decision

Deontic:

21. Advice
22. Permission
23. Request
24. Exhortation
25. Admonition
26. Instruction

F₃ - Intonational focus

1. Presentational
2. Corrective
3. Counter-presuppositional
4. Definitional
5. Contingency
6. Reactivating
7. Identificational
8. Non-focused

F₄ - Rhetorical form

1. Irony
2. Aposiopesis
3. Prepetition
4. Anacoluthon
5. Expeditio
6. Eutrepismus
7. Dialysis
8. Sentential Adverb
9. Polysyndeton
10. Asyndeton
11. Rhetorical question
12. Parenthesis
13. Epizeuxis
14. Enumeratio
15. Neutral

F₅ - Interpersonal Motivational State

1. Attachment
2. Caregiving
3. Rank
4. Sexual
5. Peer cooperation
6. Neutral

F₆ - Speech mood

1. Whispered
2. Normal
3. Shouted

F₇ - Spontaneity

1. Spoken
2. Read
3. Recited

F₈ - Punctuation form

1. Single commas
2. Null

F₉ - Emotion

1. Anger
2. Sadness
3. Fear
4. Joy
5. Love
6. Surprise
7. Disgust
8. Shame
9. Neutral

F₁₀ - Subjective expressiveness skills

(Open set)

1. Able-bodied speaker
2. Verbal dyspraxia
3. Aphasia
4. ...

F₁₁ - Social context

(Open set)

1. Thesis dissertation
2. Politic rally
3. Religious ritual
4. ...

F₁₂ - Language, dialect or local accent form

(Open set)

1. Standard Italian
2. Genoese variety of the Ligurian dialect
3. Standard Cantonese
4. ...