Collecting Data to Study Prosodic Patterns and Their Mappings to Meanings Across Languages

Abstract

In this paper we describe "AC" *: a web-based interactive game designed to collect massively multi-speaker, multi-lingual oral data on the connection between prosody and various aspects of meaning. Game participants take on the two roles of *auditioners* and *casting directors*. Auditioners are asked to record certain target phrases modulated according to the emotional or attitudinal profiles that correspond to contexts or stage cues given to them. They then switch roles and become Casting Directors. Now they have to listen to other participants' recordings, guess the corresponding context/stage cue that the auditioner tried to convey, and evaluate how good the performance was. By having the players alternate between these two roles we obtain both data creation and data validation from the same set of participants. We expect that the final dataset of labeled recordings will be valuable for a range of applications: training multilingual SER classifiers; discovering correlations and variations in prosodic patterns among unrelated languages; examining correlations between prosodic patterns and emotion recognizability; probing the possibility that some prosodic patterns are universal.

*For the sake of anonymity we abbreviate the name of the game.

Keywords: Prosody, GWAP, NLP

1. Introduction

Prosody is a fundamental aspect of spoken language, associated to units larger than the segment and related to word/sentence accent, intonation, lexical tone and rhythm/tempo (Cole, 2014). Modification of prosodic patterns is a tool we use either inadvertently or intentionally (e.g. in acting) to convey our attitudes and emotions. By changing nuances in pitch, amplitude or speed, we can remove syntactic or semantic ambiguity, alter or enhance the meaning of words, or change focus.

The study of Affective Prosody, an umbrella term that includes emotional and attitudinal prosody (Mitchell & Ross, 2013) involves mapping prosodic patterns to the information structure encoded in a text. Theoretical researchers working to establish such correspondences could benefit from a labeled set of cross-linguistic data where linguistic contexts are mapped onto utterances produced according to prosodic patterns appropriate for that context. At a computational level, data of this sort would cater to the growing interest in spoken dialogue with AI agents, which will ultimately be expected to detect emotions and attitudes in human speech and answer with suitably nuanced intonations.

Our attempt to build such a dataset takes its moves from the success of other linguistic games-with-a-purpose (GWAPs, see Ahn, 2006) like Phrase Detective (Chamberlain et al, 2008) and tries to leverage the power of competitive gaming to enroll a large number of subjects in a game of acting, inspired by the Stanislavski's method (as described in Jakobson, 1960). The byproduct of the game is the collection of highly controlled prosodic data, which are cross-validated by the very same players who provide their voices for data production.

1.1 Background

There exist various data sets that may be compared to ours, mostly designed for Speech Emotion Recognition (SER, Swain, 2018). Some contain manually annotated video recordings from Youtube labeled with the 6 basic emotions in 4 languages (CMU-MOSEAS, Bagher Zadeh, et al 2020). Others draw data from talk-shows (e.g. for German, Vera Am Tag, Grimm et al., 2008) or use a much more limited number of professional actors, instructed to record pre-set English phrases (RAVDESS, Livingstone & Russo, 2018). Without doubting the importance of naturalistic data, we believe that there is a need for a more controlled set of data where the very same linguistic expression is uttered with very different communicative intents, creating a set of expressions that differ *only* in prosody, not in the choice of words. This is difficult to obtain from naturalistic input but allows a fine-grained control of intonation parameters. The datasets above cannot provide quantitative information on the extent to which the communicative intent is met (i.e. how many listeners could recognize a given utterance as expressing e.g. *anger* and not *disgust* or *fear*), cannot be easily extended to other languages and do not address "attitudinal" uses of prosody (association with focus, syntactic or lexical disambiguation, irony, etc.). Some of these issues have been studied by phoneticians with lab experiments (for Romance languages see Origlia, et al 2014, Bocci 2013, Gili Fivela, et al. 2015) but with a limited number of speakers.

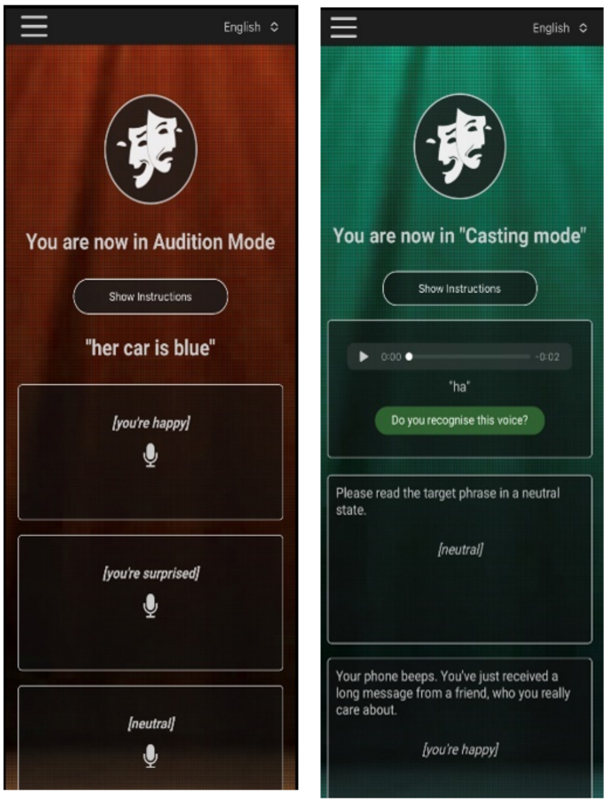
There is at least one multilingual, massively multisubject database of spoken language (commonvoice.mozilla.org), but communicative intents are not labeled and may only be inferred from the text. All things considered, the research community is still missing a comprehensive multilingual dataset containing labeled recordings from a large number of speakers, suitable for studying emotional as well as attitudinal prosodic patterns in a comparative manner across languages.

2. Our Project

This need prompted us to implement a web-based game, "AC" designed to collect large amounts of recordings expressing prosody-meaning mappings in multiple languages. These recordings are generated and validated by the same players.

2.1 Game Setup

The game works as follows: to address emotional prosody, we prepare a series of linguistic expressions (**targets**) that could be uttered in various contexts and are as neutral as possible in their affective value, such as “It’s a cappuccino”. Similarly, for the attitudinal prosody, target phrases are chosen that can lend themselves to various topics of study, e.g., focus: “Kevin isn't drinking because he is unhappy.” Next, we prepare a set of discriminating contexts in which the target phrase could be found. These contexts give the background to understand how the target should be uttered, evoking certain emotions (e.g., sad, angry, happy) or giving cues that resolve the target’s ambiguity. Example contexts for the target phrases might be: “She had asked you twice ‘Did you say coffee or cappuccino?’ and you patiently told her ‘Coffee’. Now she is handing you your cup and you almost yell ‘______’!”; “Kevin drinks a lot. No matter if he is happy, or sad. He is a classic alcoholic.” We are also experimenting with contexts that are simply stage directions (“[you are angry]”), to test the difference between linguistic and metalinguistic cues.


You are now in Audition Mode

"her car is blue"

You are now in "Casting mode"

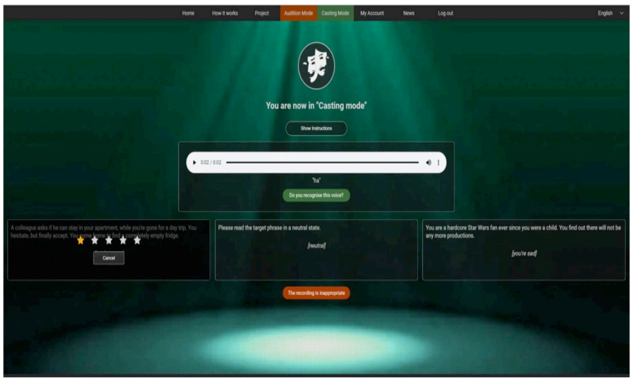
Please read the target phrase in a neutral state.

Your phone beeps. You've just received a long message from a friend, who you really care about.

Figure 1: The two modes of the AC Game, in its mobile version. Left: audition; right: casting

When they sign up, new participants are asked to provide basic information such as age, gender, region of provenance, native language and language they want to play in. They are also asked to sign an informed consent declaration, which clearly states that their anonymized data will be made publicly available under a Creative Commons Attribution Sharealike 4.0 license. Entering the game, players alternate between two roles: actors who are doing an audition (“Auditioners”) and “casting directors”, who have to evaluate actors’ performances.

- *Auditioners:* In this role participants are asked to read a randomly assigned context or stage cue and *act out* the target phrase in a way that would be best suited for the situation described in that particular context. They can record their voice multiple times, listen to their performance, submit the recording once they are satisfied and turn to the next context for the same target (there are between 2 and 4 contexts per target; see Fig.1, left). After recording the targets in all the contexts proposed, the players move on to another audition session. After a few auditions, however, they are automatically sent to the *Casting* mode.
- *Casting Directors:* in this role the player is prompted to listen to other actors’ utterances and evaluate them. Specifically, the player hears the recording of a target phrase uttered by another actor and sees the set of contexts which were presented to the actor (Fig.1, right). The task is now to assign the performance to the context for which it was intended. After matching the recording, the casting director rates the performance of the actor on a 1-5 Likert scale. (See Fig.2)


You are now in "Casting mode"

A colleague asks if he can stay in your apartment, while you're gone for a day trip. You hesitate, but finally accept. You come home to find a completely empty fridge.

Please read the target phrase in a neutral state.

You are a hardcore Star Wars fan ever since you were a child. You find out there will not be any more productions.

Figure 2: Rating the performance in Casting mode.

In the Casting mode players are also requested to filter out audio clips that have a poor sound quality, do not match the target phrase written on the page, contain inappropriate contents or somehow include tips that help the casting director in identifying the corresponding context. Such reports will be relayed to the players, who will be warned or, in more serious cases, removed from the game. An additional button allows casting directors to identify voices already heard. This can provide training data for voice recognition tasks (see AL-Shakarchy, et al., 2022).

2.2 Scoring System

The game is designed to appeal to people who would like to test their skill level at acting or recognizing the communicative intent conveyed by human voice. Each player receives scores based on how well they performed as actors, calculated as the number of times their performance was correctly matched by the casting directors, multiplied by the number of alternatives the target had (from 2 to 4). Players also receive casting points, calculated as a function of the number of times a recording meant for a given context is attributed to that context. The two scores are combined to give an overall score and a place in a public Scoreboard. Players receive score-related titles (from “Grinding stone” to “Acting god”).

2.3 Avoiding caricatures, removing abuse

One possible drawback of the AC design is that, based as it is on discrimination, it might lead to non-natural, exaggerated utterances. For instance, if all I have to do is to say “tonight” as a question or an assertion, I might simply exaggerate the raising intonation in the question, creating an unnatural, ‘caricature’ question. In other terms, focusing on context discriminability rather than prosodic appropriateness makes the actors adapt their intonation only to the specific set of contexts, as it might happen for the target in the two set (1) (worrying/non-worrying) and (2) (worrying/scary).

(1) Target: Who are you?

a. Context 1: it’s late at night and you are alone in the office. Someone knocks at the door, but you do not expect anyone. You open. It is big man, with a scar and a strange smile.
b. Context 2: it’s late at night and you are alone in the office. Someone knocks at the door. A young girl with a sweet smile stands there, a little embarrassed.

(2) Target: Who are you?

a. Context 1: it’s late at night and you are alone in the office. Someone knocks at the door, but you do not expect anyone. You open. It is big man, with a scar and a strange smile. = (1-a)
b. Context 2: it’s late at night and you are alone in the office. Someone knocks at the door. It’s a green, humanoid monster with a mouth, filled with sharp teeth.

At the data-gathering level, the presence of caricatural intonation could sometimes be a feature, not a bug, as it might be used to better highlight prosodic differences. However, it would certainly be inappropriate for other uses of the data (AI model training). To contain the damage, we employ the following features:

- Using the Performance score: beside assigning the utterance to a context, the evaluator assigns a score to it. With appropriate instructions (“Rate how natural the utterance sounds in this context”) this can be used to penalize caricatural answers. The auditioners are made aware of the fact that the rating is part of their scores.
- A high number of alternative contexts (currently 4) should make the problem less pronounced, since with many contexts it would be too difficult to contrastively tailor intonations.

Another possibility to explore is to tell the performer that at evaluation time multiple performances assigned to the same context in different auditions will be randomized. In other terms, the evaluator might be given the context set in (2), but the utterance to evaluate could sometimes be the one the actor has associated to (1-a), rather than (2-a).

3. Current Data Prompts

The game can currently be played in English, Italian, German and French, soon to be followed by Russian, Farsi and Arabic. The current version contains 35 target phrases, and 185 unique contexts. The current set of targets and contexts has been primarily designed to study the linguistic expression of the basic emotions (*fear, joy, surprise, disgust, sadness, anger*, plus *neutral,* see Ekman, 1999). We chose targets of two sizes: long and short (monosyllabic); among the short ones we tested a number of non-linguistic vocables (“Oh”, “Ah”, “Ha” for English and equivalent for the other languages). Targets were chosen to be similar and sometimes identical in the various languages. Emotions can be prompted by textual contexts (see an example in Sec. 2.1), stage directions or both, to test which cue is more effective.

Apart from emotions, current data include target/contexts to test syntactic ambiguity (PP attachment), broader attitudes (e.g. embarrassment, concern, pessimism, dignity, fake cordiality, perplexity, correction, sarcasm, grieving, boredom, pleasant and unpleasant surprise), association of negation with focus, normal vs. rhetorical questions, long distance Wh-extractions (“When did you say that John left?”). Other examples are specific for different languages (e.g. the definite vs. kind-denoting reading of the definite article in Italian; the universal vs. existential reading of bare plurals in English). Note that for some of these phenomena the system could simply provide evidence that prosody cannot disambiguate them.

Yet additional materials are fillers, designed only to make the game more fun and engaging. Some of these prompts are next to impossible and might not be part of the data released. Note however that adding new data is quite straightforward. We are presently in contact with phoneticians at the university of Padua, Siena and Bolzano and we welcome new collaborations.

4. Potential Applications

The AC data can be used to investigate a range of topics across theoretical and computational linguistics.

- Speech Emotion Recognition (SER) with a broader range of speakers from different regions and age groups. Artificial Emotional Speech Production can also be tested by injecting artificially produced samples in the Casting phase.

- Examining the effect of combining multiple intonational patterns (e.g. *question+surprise, question+emotion, multiple emotions*). The compositionality of emotions is currently mostly focused on bodily/facial features (Cavicchio, et al., 2018) but the combination of emotions in speech could benefit from a data set such as the one we are creating with AC.
- Examining how the intonation patterns vary from speaker to speaker and from language to language. Inter-speaker variation is actively studied in labs (Niebuhr, et al., 2011; Myberg, 2013; Feldhausen, 2016) but not with the large volume of data that a web game could gather. Interlinguistic variation has been discussed (see Rabanus, 2003, Gili Fivela et al. 2015), but not systematically evaluated for very different languages.
- Discovering ambiguous intonational patterns (i.e. targets consistently assigned to multiple contexts) and ordering semantic/emotional contexts w.r.t. how hard it is to consistently translate them into unambiguous prosody.
- Discovering the individual extent to which passive prosodic competence differs from active one (can one be good casting director without being a good actor, or vice versa?) and whether the recognition ability is affected by regional, gender or age differences.

5. Universality of Prosody

There seems to be a consensus that having emotions is universal among humans (Ekman et al, 1972). But is the same notion true when it comes to expressing our emotions, particularly in speech? In other words, are there any prosodic patterns that correspond to the same emotion in more than one language?

Examining this question could entail the following two approaches:

5.1. Non-Computational

One way to test whether prosodic patterns could be linked to certain emotions independent of language would be having native speakers of Lang1, listen to and label recordings done in Lang2, which is unfamiliar to them. Shakuf et al (2022) following this method demonstrated that native speakers of German and Hebrew could for the most part correctly identify emotions in the language they were not familiar with.

In AC, we have a set of stimuli where participants have to play in an invented language.

- Target phrase “Sotaki”
- Context: The woman came from a distant tribe and when she spoke, we couldn't understand her language. She had no idea what gunpowder was: when we lit up some firecrackers, she went pale in her face and said: "___"

Successful identification of emotions in this task could be interpreted as the notion that prosodic patterns can be linked to emotions independent of the language. At the moment we do not have enough data to arrive at conclusive results.

5.2. Computational

Emotional Speech Classifiers are typically trained on large datasets of recordings in a language and then tested on that same language. There have been studies in which the classifier is trained on one language and tested on another. Wish et al (2021) tested a classifier trained on Urdu on several European languages. Their results show that the model was able to correctly categorize the emotions in languages it was not familiar with.

6. Current State of The Project and Preliminary Results

The website was officially launched in the late summer of 2023, and we are at the moment promoting it via direct contacts and Italian social media. The site has at the moment 102 registered users, of which 44 are English, 43 Italian, 10 German, and 5 French. We have a total of 778 recordings.

To begin investigating the inter-speaker variation in recognition ability (3rd question above) we looked at the rate at which casters recognized the intentions of actors with the same or the opposite gender (see Table 1).

Casting Directors ↓	Actors	
	Male	Female
Male	42,1	57,8
Female	54,8	45,1

Table 1: Correct recognition %, by gender

This preliminary results hint at female emotions being better recognized than male, and at an advantage in recognizing emotions in the opposite gender.

To test the effectiveness of the data collected for SER, we trained a classifier, (Hubert, Hsu, 2021) on the English utterances recorded by AC players and tested the model on the Italian dataset. Our results, again very preliminary, show that even without cleaning the files available so far the English model reaches 43% accuracy on Italian (chance = 25%), showing that prosodic patterns do transfer from English to Italian.

7. Bibliographical References

Ahn, L. v. (2006). Games with a purpose. *Computer*, 92-94.

AL-Shakarchy, N. D., Obayes, H. K., & Abdullah, Z. N. (2022). Person identification based on voice biometric using deep neural network. *International Journal of Information Technology*, 789–795.

Bagher Zadeh, A., Cao, Y., Liang, P., Poria, S., & Morency, L.-P. (2020). CMU-MOSEAS: A Multimodal Language Dataset for Spanish, Portuguese, German and French. *EMNLP*, (pp. 1801-1812).

Bocci, G. (2013). The Syntax Prosody Interface: A cartographic perspective with evidence from Italian. Amsterdam/Philiadephia: John Benjamins Publishing Company.

Cavicchio, F., Dachkovsky, S., Leemor, L., Shamay Tsoori, S., & Sandler, W. (2018). Compositionality in the language of emotion. *PloS one*.

Chamberlain, J., Poesio, M., & U., K. (2008). Phrase Detectives: A web-based collaborative annotation. *International Conference on Semantic Systems.*

Cole, J. (2014). Prosody in context: a review. *Language, Cognition and Neuroscience*, 1-31.

Ekman, P. (1999). *Handbook of cognition and emotion.*

Ekman P., Friesen W. V., Ellsworth P. (1972). *Emotion in the human face: Guidelines for research and an integration of findings*. New York, NY: Pergamon Press.

Feldhausen, I. (2016). Inter-speaker variation, optimality theory, and the prosody of clitic left dislocations in Spanish . *Probus*.

Gili Fivela, B., Avesani, C., Barone, M., Bocci, G., Crocco, C., D'Imperio, M., & Sorianello, P. (2015). Intonational phonology of the regional varieties of Italian. Dans P. P. Sónia Frota (ed.), *Intonation in Romance* (pp. 140–197). Oxford University Press.

Grimm, M. K. (2008). The Vera Am Mittag German Audio-Visual Emotional Speech Database . *IEEE international conference on multimedia and expo*, (pp. 865-868).

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *Computer Science*.

Jakobson, R. (1960). Linguistics and Poetics. Dans T. Sebeok, *Style in Language* (pp. 350-377). Cambridge: Massachusetts Institute of Technology Press.

Mitchell, R. L., & Ross, E. D. (2013). Attitudinal prosody: What we know and directions for future study. *Neuroscience & Biobehavioral Reviews*, 471-479.

Myberg, S. (2013). Sisterhood in Prosodic Branching *Phonology*.

Niebuhr, O. D., Fivela, B. G., & Gangemi, F. (2011). Are there "shapers" and "aligners"? individual differences in signalling pitch accent category. *17th ICPhS*, 120–123.

Origlia, A., Cutugno, F., & Galatà, V. (2014). Continuous emotion recognition with phonetic syllables. *Speech Communication*, 155-169.

Rabanus, S. (2003). A Cross-Linguistic Study of German and Italian. *Zeitschrift für Sprachwissenschaft*.

Shakuf, V., Ben-David, B., Wegner, T. G., Wesseling, P. B., Mentzel, M., Defren, S., ... & Lachmann, T. (2022). Processing emotional prosody in a foreign language: the case of German and Hebrew. Journal of Cultural Cognitive Science, 6(3), 251-268.

Steven R. Livingstone, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE*.

Swain, M. R. (2018). Databases, features and classifiers for speech emotion recognition: a review. . *Int J Speech Technol* , 93–120.

Wisha, Z., Javed, A., Khan, J., & Ghadekallu, T. (2021). Cross corpus multi-lingual speech emotion recognition using ensemble learning. *Complex & Intelligent Systems*, 1-10.