The "Actors Challenge" Project

Collecting data on intonation profiles via a web game

Natallia Chaiko, Sia Vosh Sepanta and Roberto Zamparelli
CIMEC, University of Trento
name.surname@unitn.it

Abstract

This paper describes "Actors Challenge", a soon-to-go-public web game where the players alternate in the double role of actors and judges of other players' acted-out utterances, and in the process create an oral data set of prosodic contours that can disambiguate textually identical utterances in different contexts. The game is undergoing alpha testing and should be deployed within a few months. We discuss the need, the core mechanism and the challenges ahead.

Keywords: GWAP, prosody, Web games, NLP

1. Introduction

The study of intonation is an important part of semantic research, as it affects information structure, speaker's attitudes, structural ambiguity resolution and other syntactic and semantic phenomena. While there are now various well-established ways to annotate prosodic features (Pierrehumbert and Hirschberg, 1990; Gussenhoven, 2002) and tools to facilitate the annotation (see the recent ProsoBeast developed by (Gerazov and Wagner, 2021)), the exact mapping between prosodic features and semantics is not a solved problem, as is the consistency of such mapping across speakers and languages. While interesting attempts at a compositional theory of meaning/intonation have been done (see especially (Steedman, 2014; Schlöder and Lascarides, 2020)), they appear to be fairly language-specific, and do not consider the interaction between information structure and emotion. Similarly, some of the current research on the left-periphery of the sentence (devoted to (contrastive) topics and focus, question intonation, etc., e.g. (Frascarelli, 2010; Bianchi and Frascarelli, 2010)) rely on subtle prosodic cues which have not been verified by large pools of speakers, and whose consistency may be difficult to evaluate.

On another front, the study of emotions has been increasingly gaining attention due to its direct application to AI. In particular, comparative research across languages and cultures in word meanings, among them emotion words, has revealed interesting results and common patterns (see e.g. (Thompson et al., 2020)). Consequently, interest in data sets that revolve around emotions in speech has been steadily on the rise. One of the most recent ones, multilingual as well as multimodal, is the CMU-MOSEAS data set with over 40K labeled sentences (Bagher Zadeh et al., 2020). Once again, although the utterances are labeled according to the type of emotion they try to convey, the prosodic patterns are not annotated.

All of this research could profit from a large, multilingual, multi-speaker data set which reliably associates intonations and meanings in a controlled set of cases. To the best of our knowledge, a data set of this sort does not yet exists. The project closest to the one described in this proposal is the Mozilla-funded project Common Voice (Ardila et al., 2020), where volunteer speakers read sentences in their own languages and evaluate sentences read by others. The data set thus collected has broad language coverage (76 languages) and many hours of speech (about 2K validated hours just for English). However, sentences are presented and evaluated out of context, so there is no mapping between intonation and semantics beyond what can be extracted from the short passage to be read. A single sentence may be read differently by different speakers, but these differences cannot be traced to different discourse-level effects associated with them, or to the emotions the speaker intended to convey. There is also no incentive for careful validation, and no check to make sure that sentences are validated by speakers of the same variety, or even the same language.

2. Proposal

To address these gaps, and building on the experience gained from the oral data collection project VinKo, we propose a social web game, Actors Challenge (AC), designed to collect and validate large amounts of data on the correspondence between the intonation of a linguistic expression and its meaning in context. The success of projects such as DALI, on anaphora annotation (Poesio et al., 2013) and other linguistic data collection (Kicikoglu et al., 2020), has convinced us that intonation is a domain that could be appropriate for a 'serious game' design, administered over the web and mobile-friendly. This would also make it easy to deploy the game in multiple languages, so as to collect data comparable with materials from more traditional oral data repositories (e.g. VoxForge). Our plan is to initially launch the website interface and contents in English and Italian. After analyzing the pattern of usage and refining the materials, we plan to add German and Farsi, with

ultimate goal of seeking out the collaboration of linguists abroad and expand the project to various other languages.

The basic setup (which draws from a method attributed to the Stanislavski's acting method by Roman Jakobson) (Jakobson, 1960) runs as follows.

- The researcher produces a linguistic expression, the *target*, which should be chosen to be very general (i.e. something that could be uttered in many different contexts, like *good evening*, *that's right*) and to be phonetically well distinguishable (to facilitate spectrographic analysis). Ideally, the target should be text that can also be easily adapted to different languages.
- The researchers then create a set of textual *discriminating contexts* in which the target could be uttered. Contexts, which could precede or follow the target, could be either the sentences adjacent to it (*John, boring Mary to death??* target = *Bill spoke to her for the whole evening*), or explicit descriptions of the circumstances in which it should be uttered ("You have just discovered a thief under your bed, and you say..." target = *Good evening*) or just bare stage directions (e.g. [*with affectation / with bitterness / pensive*]).
- The target's contexts give the background to understand how the target should be uttered, setting up the informational focus of the utterance, providing contrast or triggering different intonational profiles on the basis of their emotional content (i.e. surprise, fear, disgust, hurry, affection, hesitation, irony, etc.).

On the gaming side, the web site aims to attract players by offering them the opportunity to challenge each other on their 'active' and 'passive' acting skills: how much meaning and expression they can convey with their voice alone, and how fine-grained their understanding of other players' vocal nuances is. The mechanism works as follows.

The players log in into a web site, fill out a questionnaire (language and variety they identify with, gender, age) and are assigned to one of two roles: *audition* or *evaluation (casting)*. In the first one, they play the role of an actor that auditions for a part; in the second, they evaluate other players' performances and decide whether they correspond to a given context; the entire process is anonymous both ways. More specifically:

- In the *audition mode* the participant sees a (randomly chosen) written target sentence and a set of text-boxes containing the contexts (see Fig. 1). The participant selects one of the contexts, and records his/her voice uttering the target, aiming to implement the intonation that he or she feels appropriate for the context selected. The participant can listen to his/her recording, verify voice and recording quality, approve it if satisfied or repeat the recording. The auditioner then selects a different context, and repeats until all the contexts have a recording. The target is the same for all the contexts so, crucially, only the prosody can distinguish one from the other.

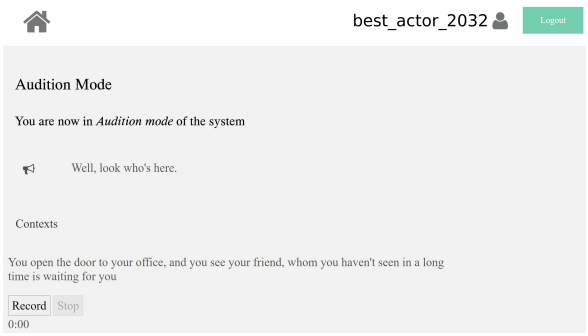

Figure 1: Detail of an Audition page screenshot.

- In the *evaluation mode*, the participant moves to a page with a set of contexts and a single loudspeaker icon (see Fig 2). Clicking on the icon, the evaluator hears a target that has been recorded by another participant in the auditioner role, and sees the set of contexts that was presented to the actor (in random order). The task is now to guess the context for which that intonation was meant. After the choice is done, the evaluator assigns a score to 'how convincing/natural' the performance was for the context chosen, using a 1–5 Likert scale (*performance rating*).

  A "Signal abuses" button is available at this stage to remove audios that have low sound quality, do not match the intended sentence, contain inappropriate contents, or add cues to facilitate context identification. These info can be used to alert the player and can trigger removal of the utterance and/or player.
- After a certain number of trials in one role participants are forced to do the other role, so as to maintain a balance.
- The primary measure of how good that intonational profile was at discriminating the semantics provided by the contexts — and thus how good an actor its utterer was — is the success rate of the evaluators in matching the recorded target with the context intended by the person who uttered it. A secondary measure is the 1-5 rating given to the performance by the evaluators. This is considered only if the attribution of the target to a context matches the intended context. Suppose, for instance, that player alpha had to utter "That's good." in four contexts A, B, C, and D. The player's utterance for context B is sent to 10 evaluators, 7 of which correctly assign it to context B, 1 to context A and 2 to C. The average rat-

ing assigned to alpha's utterance by the 7 evaluators who correctly classified it as meant for B contributes to alpha's score, along with the 7-out-of-10 proportion. The final score is given by the results for each of alpha's utterances (i.e. also those meant for contexts A, C and D).

- Players are also scored in their role as evaluators (the 'passive' side of acting). In this case the score is given by the variance of their judgments with respect to other evaluators' judgments on a set of cases for which there is a high level of correct identification. Scoring the evaluator's role should help increase the players' motivation in a task that could be perceived as less entertaining.

  When the performances of the participants have been judged by a sufficient number of evaluators, their *acting* and *evaluator scores* gets posted on a scoreboard. The players then receive an email from the system with an invitation to check their scores on the website and play again in the challenge.

- The acting scores will be organized in tiers, each linked to the names of famous actors. We will consider implementing the idea, suggested to us by an anonymous reviewer, that the acting score drops with time when left unused, as well as the possibility that advanced players gain the possibility of suggesting new contexts and targets for others to play. Taken together, these measures should motivate the players to return regularly to the site.

- From the researcher's viewpoint, intonation patterns which are consistently matched to a certain context and which have good ratings count as *validated data*: sound files with intonations which express a certain semantic content. The researcher also receives *negative* data: which intonation patterns are systematically associated to the wrong context, and which semantic contexts systematically fail to be disambiguated by intonations.

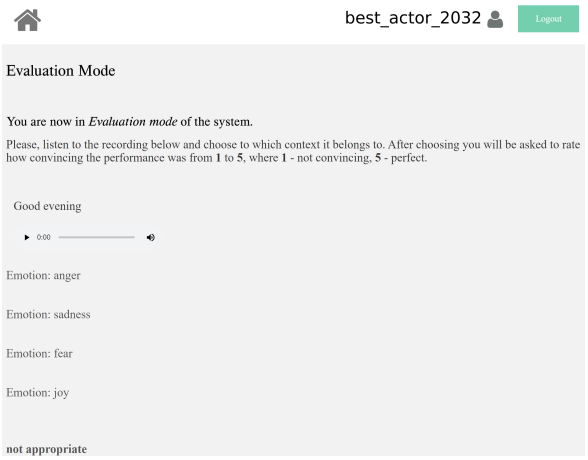

Figure 2: Detail of an Evaluation page screenshot.

3. Research targets

The outcome of the collection process described above would be a large set of intonations for the same linguistic expressions, along with the context or contexts to which they have been consistently associated (possibly, this could be distilled into a set of semantic features associated with that context, derived via crowdsourcing or via distributional semantic techniques). This material can be used for a variety of purposes, some of which require the possibility of automatic phonetic analyses of large amounts of data (but data with largely invariant lexical content). With the help of colleagues with an expertise in prosody and the meaning-intonation interface, we intend to look at the following topics.

- Examining the effect of combining multiple intonational patterns (e.g. question+surprise, question+emotion, multiple emotions). The compositionality of emotions is currenly an active research topic, but is mostly focused on bodily/facial features (see e.g. (Cavicchio et al., 2018)). The combination of emotions in speech, on the other hand, is an area that is relatively new and could benefit from a data set such as the one created by our AC project.
- Speech Emotion Recognition (SER): In the past 10 years the CL community has been busy developing models that would recognize emotion in spoken language (see (Yoon et al., 2018)); an essential factor in the effectiveness of these models is the data they are being trained on. We believe AC could contribute to building up this corpus.
- Examining how the intonation patterns varies from speaker to speaker. Inter-speaker variation is actively studied in labs (Niebuhr et al., 2011; Myrberg, 2013; Feldhausen, 2016) but not with the large volume of data that a web game could be expected to gather. Aspect to consider for investigation include irony, the difference between rhetorical and non-rhetorical questions, the theme/rheme distinction and the resolution of structural ambiguity. The amount of data allowed by a GWAP approach could also make possible to study the interactions among these phenomena.
- Examining how the intonation patterns vary across languages for the same semantic cues (translations of the same targets/contexts)
- Discovering ambiguous intonational patterns (i.e. targets consistently assigned to multiple contexts) and ordering semantic/emotion context 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 (i.e. to what extent one can be a good evaluator without being a good actor and vice versa)

- Testing to what extent evaluators can correctly classify performances from actors from different parts of the country, and possibly even different languages. Note that normally evaluators will be asked to evaluate the performances of people in the same area, obviously excluding one's own performance.
- Probing the ability of players to recognize other players' individual *voices*. This data will be gathered by adding a yes/no question to the evaluation mode: "Do you think you have heard the voice of this actor before?". Comparing the answer to the history of auditioners the player has encountered gives us the ground truth.

Beyond this specific research questions, we believe that the data collected with a game, if successful, can be extremely valuable to training general computational models of intonation, both in production and in perception. All the data collected, anonymized in conformity to the EU GDPR policy, will be made available to the public under a Creative Commons BY-SA 4.0 license.
Last but not least, we will explore the idea of using this data as an ingredient in the creation of distributional multi-modal meaning representations of emotion terms, associating e.g. "fear" to the set of intonations that people use to render fear contexts.
.

4. Avoiding caricatures, removing abuse

One possible drawback of the Actors Challenge 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/nonworrying) 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 large toothed mouth.

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 plan to 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 higher 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).

As in any distributed data gathering exercise, our game presents a trade-off between sound quality (with poor recordings due to low quality microphones, noise, speaker's volume or other factors) and amount of data (Lafourcade et al., 2015). The possibility for the actors to listen back to his/her own utterance before submitting it should partly address this, as could the shared experience as evaluators, which would raise the participant's awareness of the importance of good sound. Using the game on mobile devices could also help, since cellphones' microphones are often better than PC's and the actors are likely to speak closer to the mike; noise will worsen going mobile, but there are good tools to clean up this aspect at data-preparation time. Evaluators have a button to raise alarm about the poor sound quality of specific utterances, and repeated alerts are fed back to the players at log in.
Another concern is the possibility of abuse. This could take the form of completely inappropriate recordings (e.g. insulting remarks replacing the target) or attempts to conditions the outcome by adding information above and beyond the intonation. To counter this possibility, the evaluators have a "Signal abuse" button. Multiple abuse alerts on one utterance lead to exclusion of that utterance from further evaluation. Repeated cases lead the system to (temporarily or permanently) ban players. We will also experiment with dictation software to double check if the utterance matches the target.

5. The current state of the project and its future

The software engine for the audition/evaluation has been developed in Java by one of the authors and is ready to be deployed, modulo minor feature addition. The front-end of the website is currently under revision, with the goal of giving it a more refined, game-like look and making it suitable for mobile devices. The next step will be to adapt the new interface to the engine. After these steps, the site will be open to beta testers by summer 2022. If this phase is successful, we plan to advertise it among a limited circle of users, whose feedback will help us fine-tune the game (materials, feedback parameters, interface) and improve interactive features, like the scoreboards. We will then advertise it on social media and start the real data-gathering exercise. In parallel, we will be expanding our set of contexts and languages (currently only English and Italian), and translating the interface (currently only in English). As mentioned above, the contexts include textual descriptions of the circumstances in which the target is uttered, including emotion cues and focus. Researchers interested in using our tool could provide further targets and contexts in the form of a spreadsheet. We will however work hard to make sure that the game contains enough playful material to keep the players entertained: "serious games" should never be as serious as labs.

6. References

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., , and Weber, G. (2020). Common voice: A massively-multilingual speech corpus. In *Proceedings of LREC*.

Bagher Zadeh, A., Cao, Y., Hessner, S., Liang, P. P., Poria, S., and Morency, L.-P. (2020). CMU-MOSEAS: A multimodal language dataset for Spanish, Portuguese, German and French. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1801–1812, Online, November. Association for Computational Linguistics.

Bianchi, V. and Frascarelli, M. (2010). Is topic a root phenomenon? *Iberia: An International Journal of Theoretical Linguistics*, 2(1).

Cavicchio, F., Dachkovsky, S., Leemor, L., Shamay-Tsoory, S., and Sandler, W. (2018). Compositionality in the language of emotion. *PloS one*, 13(8):e0201970.

Feldhausen, I. (2016). Inter-speaker variation, optimality theory, and the prosody of clitic left-dislocations in spanish. *Probus*, 28(2):293–333.

Frascarelli, M. (2010). Narrow focus, clefting and predicate inversion. *Lingua*, 120(9):2121–2147.

Gerazov, B. and Wagner, M. (2021). Prosobeast prosody annotation tool.

Gussenhoven, C. (2002). Intonation and interpretation: Phonetics and phonology. In *Speech Prosody 2002, International Conference*.

Jakobson, R. (1960). Linguistics and poetics. In T. Sebeok, editor, *Style in Language*, pages 350–377. Massachusetts Institute of Technology Press, Cambridge, MA.

Kicikoglu, O. D., Bartle, R., Chamberlain, J., Paun, S., and Poesio, M. (2020). Aggregation driven progression system for GWAPs. In *Workshop on Games and Natural Language Processing*, pages 79–84, Marseille, France, May. European Language Resources Association.

Lafourcade, M., Joubert, A., and Le Brun, N. (2015). *Games with a Purpose (GWAPs)*. (Focus Series in Cognitive Science and Knowledge Management). John Wiley & Sons.

Myrberg, S. (2013). Sisterhood in prosodic branching. *Phonology*, 30(1):73–124.

Niebuhr, O., D'Imperio, M., Fivela, B. G., and Cangemi, F. (2011). Are there "shapers" and "aligners"? individual differences in signalling pitch accent category. In *17th ICPhS*, pages 120–123, Hong Kong.

Pierrehumbert, J. and Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. Cohen, et al., editors, *Intentions in Communication*, pages 271–312. MIT Press, Cambridge, Mass.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., and Ducceschi, L. (2013). Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Trans. Interact. Intell. Syst.*, 3(1):3:1–3:44, April.

Schlöder, J. J. and Lascarides, A. (2020). Understanding focus: Pitch, placement and coherence. *Semantics and Pragmatics*, 1(13).

Steedman, M. (2014). The surface-compositional semantics of english intonation. *Language*, 90(1):2–57.

Thompson, B., Roberts, S. G., and Lupyan, G. (2020). Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, 4(10):1029–1038.

Yoon, S., Byun, S., and Jung, K. (2018). Multimodal speech emotion recognition using audio and text.