POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

VoiceMood: Assessing Emozionalmente for Spontaneous Emotion Recognition in Conversational Speech

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Authors: Gabriel Bondioni, Paolo Carpinteri

Advisor: Prof.ssa Franca Garzotto

Co-advisor: Dott. Francesco Piferi

Academic year: 2024-2025

1. Introduction

Human emotions are central to communication, modeling both how individuals express themselves and how messages are interpreted by others. In recent years, the field of *Affective Computing*, and more specifically *Speech Emotion Recognition (SER)*, has attracted significant attention. SER aims to enable machines to recognize and respond to emotional signals in speech, with applications in healthcare, education, human–computer interaction, and team building.

Despite these advances, important challenges remain. Existing emotional speech corpora often rely on acted emotions, scripted sentences, and controlled laboratory conditions [4, 5]. While these datasets are useful for benchmarking, they fail to capture the variability and authenticity of everyday communication. In particular, Italian corpora remain scarce and limited to short, controlled recordings such as EMOVO [2], DEMoS [3], and Emozionalmente [1]. EMOVO is based on acted speech, DEMoS introduced some induced emotions, while Emozionalmente exploited crowdsourcing with simulated recordings, achieving encouraging results. However, all these corpora present limitations in spontaneity and recording length.

This thesis addresses these gaps with the design and development of **VoiceMood**, a web-based application for the collection of spontaneous, conversational emotional speech in Italian. VoiceMood integrates gamification and a multi-perspective annotation scheme to build a dataset closer to real communication.

2. Objectives and Research Questions

The work followed four main objectives:

1. **Design and implement VoiceMood**: create a fully functional web application for conversational emotional speech collection.
2. **Build a novel dataset**: collect spontaneous and more realistic Italian emotional speech.
3. **Fine-tune an existing AI model**: test the usability of collected data for model improvement.
4. **Evaluate and compare**: assess results against existing corpora such as Emozionalmente.

From these objectives, the following research questions guided the thesis:

- How can spontaneous conversational emo-

tional speech in Italian be effectively collected through a web-based framework?
- How do gamification strategies (competitive guessing, feedback, leaderboards) affect participation and data quality?
- How does human recognition of emotions compare with machine recognition across conditions (recording length, spontaneity, recording environment)?
- How do dataset characteristics, such as recording's length, recording environment, and speaker type, impact recognition accuracy?
- Can a triangulated annotation scheme (speaker intention, partner perception, AI prediction) reveal mismatches useful for improving labels and training models?

To answer these questions, we implemented a live voice message–based chat in which users discuss predefined topics while expressing the emotions these topics elicit. To increase engagement, we integrated an emotion-guessing game: after each exchange, users attempt to guess their partner's expressed emotion, and the AI model performs the same task. This gamified approach was designed to motivate participants and encourage the collection of richer data. The evaluation of the recordings then enables a comparison between human guessing performance and AI recognition across different conditions. Furthermore, fine-tuning an existing AI model with the collected data and comparing its accuracy with results on existing corpora allows us to assess whether the characteristics of our approach lead to performance improvements. Finally, the use of three labels for each recording, intended, perceived, and predicted, provides insights into how human perception compares with AI recognition and helps identify emotion-specific patterns that could guide future model improvements.

3. Methodology

3.1. Design of VoiceMood

VoiceMood was designed as a chat-like application where pairs of users exchanged short voice messages on predefined **topics**. Examples include:

- Imagine winning a huge amount of money in the lottery overnight.
- Reading a paper book or listening to an audiobook: does it really make a difference?
- When a friendship ends, is it better to talk about it or let it go?
- Climate change will affect our lives more than we imagine.
- Have you ever felt like your smartphone was spying on you?

At the start of each session, participants chose a **target emotion** based on the topic. At the end, both partners and the integrated AI model guessed the expressed emotions, producing three labels per recording:

1. **Intended label**: emotion chosen by the speaker.
2. **Perceived label**: emotion guessed by the partner.
3. **Predicted label**: emotion detected by the AI model.

This *triangulated annotation* captures how emotions are expressed, perceived, and predicted, enriching the dataset with multi-layered information and also serves as a game creating a **Human vs Human** and **Human vs Machine** challenge.

3.2. Technical Implementation

The system architecture is composed of:

- **Front-end:** developed in Vue.js, providing a responsive and intuitive interface.
- **Back-end:** based on Javascript, supporting scalability and real-time communication.
- **Database:** hosted on Supabase, designed for anonymization, secure access, and efficient storage of recordings.

Privacy, scalability, and device independence were key design constraints. The application can be accessed from any device using a browser, lowering barriers to participation.

4. Data Collection

4.1. Usability Test

An initial usability test was conducted with 14 participants to evaluate the accessibility and overall quality of the VoiceMood platform. The assessment employed a questionnaire based on the *System Usability Scale (SUS)*, a widely used instrument for measuring perceived usability,

complemented with application-specific questions.
Overall, the results were encouraging, with the majority of responses being positive. This indicates that the system was perceived as intuitive and easy to use. Participants reported that the chat-like interface was clear, the gamification elements enhanced engagement, and the overall workflow was straightforward. The results of the *SUS* are showed in Table 1.

	Score	Interpretation
Mean SUS	78.5	Good
Median SUS	80.0	Excellent
% > 68	85%	Above Average

Table 1: System Usability Scale (SUS) results for the usability test.

Qualitative feedback highlighted a few areas for improvement, such as the need for features that could further motivate participants to use the application on a regular basis, even though the intended usage was limited to approximately five minutes per day.
These findings confirm that VoiceMood provides a solid foundation for large-scale data collection, demonstrating both usability and user satisfaction from the earliest testing phase.

4.2. Crowdsourced Test

Following the usability test, a larger-scale crowdsourced experiment was conducted to develop the VoiceMood corpus. A total of **159 non-professional Italian native speakers** were invited to take part in this phase, engaging in topic-guided conversational tasks designed to elicit emotional speech. The experiment lasted ten days, and the conversations addressed **five different topics**, ranging from everyday experiences to emotionally charged situations.
However, participation was lower than expected. In total, the study produced only **15 conversations**, corresponding to **29 audio recordings** and approximately 827 seconds of speech (around 14 minutes). The amount of data collected was therefore well below expectations and insufficient to fulfill the initial goal of fine-tuning an AI model with this dataset.
Causes for low adherence included:

- **Synchrony constraints**: requiring both users online simultaneously.
- **Lack of incentives**: no tangible rewards or recognition.
- **Limited campaign duration**: only ten days of availability.

Possible solutions include asynchronous interaction, gamified leaderboards, and longer campaigns.

5. Results

5.1. Analysis of our data

The analysis combined recordings from both the usability and crowdsourced tests, yielding a total of 112 audio samples with a duration of approximately 36 minutes. The data showed a predominance of joy, mainly due to the specific topic chosen for the usability test, while the recordings collected during the crowdsourcing campaign showed a more balanced distribution of emotions across topics. Contrary to expectations, neutrality was underrepresented, suggesting that the selected topics encouraged participants to speak with greater emotional involvement. In Figure 1 the number of recordings per emotion is displayed.

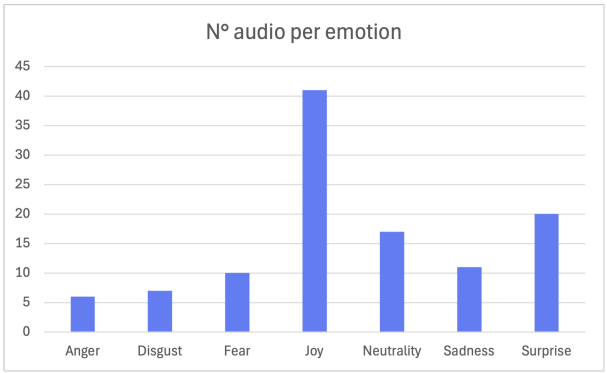

Figure 1: Graph representing the number of audio collected per emotion

Another key objective was to collect longer audio recordings compared to existing corpora. This goal was successfully achieved, as more than **80%** of the collected samples last more than ten seconds, in contrast with the shorter recordings that typically characterize resources such as *Emozionalmente* [1]. More details about that in Figure 2.

5.2. Assessing Emozionalmente

A further analysis evaluated how the pre-trained *Emozionalmente* model performed on the data

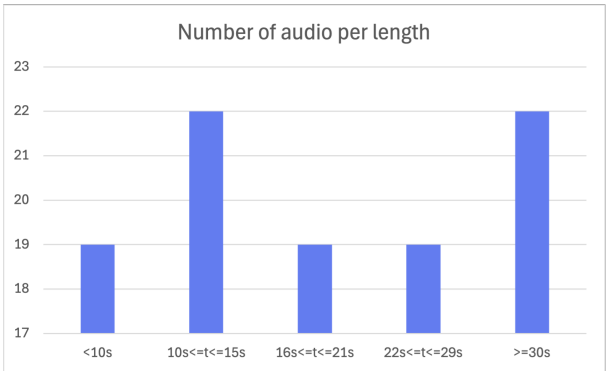

Figure 2: Graph representing the distribution of the length of the recordings

collected with VoiceMood. This was motivated by the intention to test the model in a new context and compare its results with both human performance and its previously reported accuracy.

Human participants, who attempted to guess their partner's intended emotion during the chats, achieved a recognition rate of about **62%** (see Figure 3), lower than *Emozionalmente's* accuracy [1]. By contrast, the *Emozionalmente* model reached only **12% accuracy** (see Figure 4) when considering the most probable emotion predicted for each audio. When the second most probable emotion was also included, accuracy increased to **29%** (see Figure 5), which is lower than human performance but still provides meaningful insights.

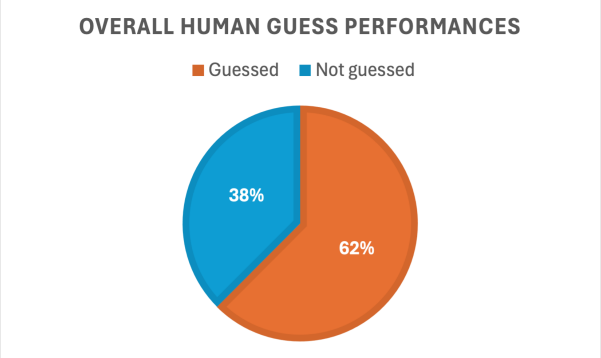

Figure 3: Graph representing human emotion guessing performances

Although these results are far from the accuracy originally reported for *Emozionalmente*, they are not completely unsatisfactory compared to human guess, especially given that the AI cannot interpret conversational context and users were able to guess on the overall chat rather than on individual audio segments.

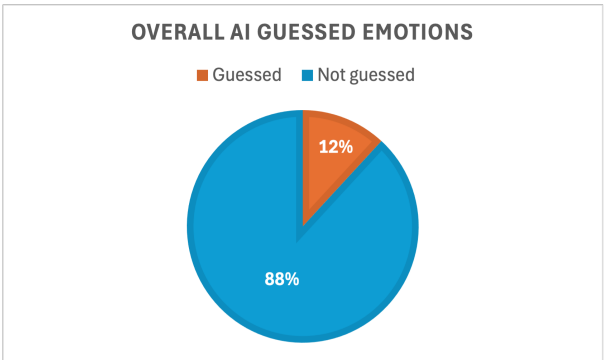

Figure 4: Graph representing overall AI guessed emotions

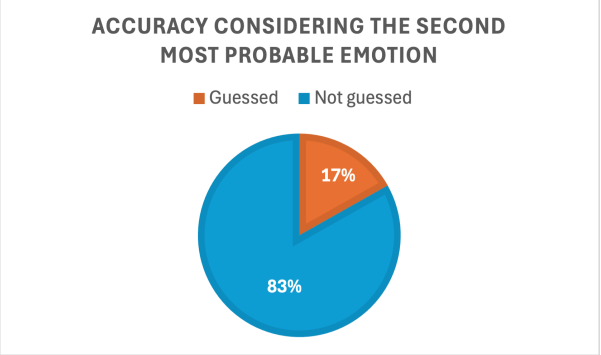

Figure 5: Graph representing AI recognition performance on second most probable emotion

6. Conclusions and Future Developments

This study demonstrated the feasibility of collecting spontaneous conversational emotional speech in Italian through a dedicated web application. The development of VoiceMood resulted in a functional platform that combines a chat-like interface, gamification strategies, and a triangulated annotation scheme to produce richer and more natural valid data compared to existing Italian corpora.

The project achieved several objectives: the implementation of a usable and engaging system, the collection of spontaneous recordings of longer duration than those typically found in acted corpora, and the first evaluation of the *Emozionalmente* model on conversational data. The results showed that, while human participants recognized emotions with moderate accuracy, the performance of *Emozionalmente* dropped significantly on our dataset. This out-

come reflects not only the limited size of the collected corpus but also the challenges posed by its lower acoustic quality compared to acted data.
At the same time, the work also faced limitations. The quantity of data gathered through the crowdsourcing campaign was lower than expected, due in part to synchrony constraints, lack of incentives, and the limited duration of the study. As a result, it was not possible to fine-tune an AI model effectively on the collected dataset. Nevertheless, these challenges provide valuable lessons for future large-scale data collection initiatives.
The originality of this work lies in its methodological contribution: the integration of gamification and a multi-perspective annotation framework in the context of Italian conversational speech. This approach highlights both the potential and the difficulty of collecting spontaneous emotional data, an essential but still underexplored resource for Speech Emotion Recognition.
Future research should focus on increasing user engagement through additional gamification elements, enabling asynchronous interaction, and developing a mobile application to lower participation barriers. With these improvements, longer and larger campaigns could generate sufficient data to train new SER models and support systematic comparisons with existing corpora.
In conclusion, while the dataset produced in this study remains limited in size, the work provides an important proof of concept. VoiceMood represents a step forward in bridging the gap between controlled laboratory datasets and the complexity of real conversational dynamics, offering both practical tools and methodological insights for the advancement of affective computing.

References

[1] Fabio Catania, Jordan Wilke, and Franca Garzotto. Emozionalmente: A crowdsourced corpus of simulated emotional speech in Italian. *IEEE Transactions on Audio, Speech and Language Processing*, pages 1–14, 01 2025.

[2] Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, and Massimiliano Todisco. EMOVO corpus: an Italian emotional speech database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 3501–3504. European Language Resources Association (ELRA), May 2014.

[3] Emilia Parada-Cabaleiro, Giovanni Costantini, Anton Batliner, Maximilian Schmitt, and Björn W. Schuller. DEMoS: an Italian emotional speech corpus. *Language Resources and Evaluation*, 54(2):341–383, 2020.

[4] Björn Schuller. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61:90–99, 04 2018.

[5] Dimitrios Ververidis and Constantine Kotropoulos. Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162–1181, 2006.