Oscar Wezenaar
o.a.h.wezenaar@student.tudelft.nl
Supervisor: Bernd Dudzik

Survey of Interrater Agreement in Automatic Affect Prediction for Speech Emotion Recognition

Introduction

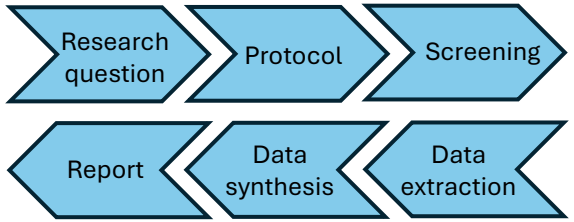
Affect refers to the subjective state of feeling [1]. **Speech Emotion Recognition (SER)** can predict emotion from speech audio. To train a SER model, **labeled datasets** are crucial. As the labeling of emotion can be subjective, an **Interrater agreement** for specific records can improve the reliability and accuracy of the dataset [2].

Research Question

"How do existing datasets for Speech Emotion Recognition differ concerning interrater agreement measures?"

Methodology

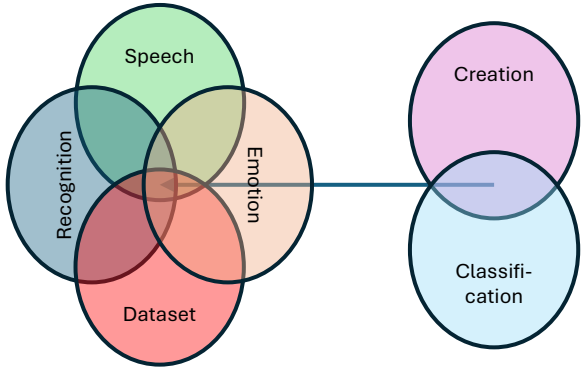
Systematic review:



There are six main steps to a systematic review as shown above.

- The protocol exists of identifying eligibility criteria, creating a search query, and creating a roadmap for data extraction.
- Papers are screened by feasibility
- Papers get screened by title, abstract, and full text.
- Data specified in the protocol is extracted and later synthesized.
- The report is written according to the PRISMA checklist [3].

Search query:



Inclusion criteria

- Paper describes the creation of a dataset.
- The paper uses a scheme or a consistent way to represent emotion
- The dataset is or can be used for SER

Exclusion criteria

- The paper is not written in English
- The dataset does not use human labeling to label emotion
- The dataset represents emotion binary.
- The paper describes a multi-modal dataset and this dataset does not store speech with emotion labels separately.
- The dataset uses gibberish as emotional speech.

Results

- 45 papers were included in the review.

Results

Affect Representation Scheme

Categorical	40	Primary emotions, Variations on Ekman's
Dimensional	9	Valence Arousal Dominance, Plutchik

Collection methods for Speech Corpora:

Actors	28
Inducing emotions	3
External sources	16

Figure 2: Frequency of Interrater agreement measures

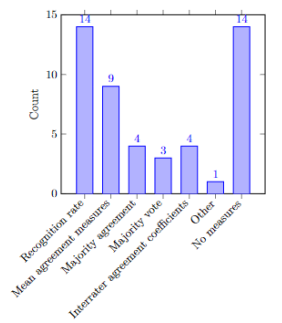
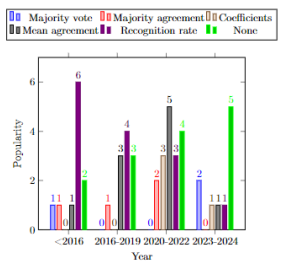


Figure 3: Popularity of Measuring Types Over the Years



- **Recognition rate** is a popular measure in datasets that use actors. However, many of these do not use interrater agreement.
- A form of **Mean agreement measures** are used in all dimensional datasets.
- **Interrater agreement coefficients**, such as kappa coefficients, are statistical measures that calculate interrater agreement.

Conclusions and Discussion

- Creating a dataset is not always the main goal of a paper.
- Language, age, and gender were not taken into account.
- Interrater agreement measures used, dependent on ARS and collection method
- Recognition rate when audio is collected with actors, mean agreement measures for dimensional datasets.
- No standard interrater agreement measure in SER datasets.

Future Work

- Research how interrater agreement affects the empirical performance of SER models.
- Research the effect of factors such as language, age, and gender on interrater agreement.
- Research the possibility of standardizing interrater agreement measures for better comparability.

References

[1]. Alpert and A. Rosen, "A semantic analysis of the various ways that the terms "affect," "emotion," and "mood" are used," Journal of Communication Disorders, 1990.

[2] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," Communications of the ACM, 2018.

[3] (2020) PRISMA 2020 checklist. [Online]. Available: https://www.prisma-statement.org/