# Analysis and Assessment of AvID: Multi-Modal Emotional Database

Rok Gajšek*, Vitomir Štruc*, Boštjan Vesnicer*, Anja Podlesek†, Luka Komidar†, and France Mihelič*

*Faculty of Electrical Engineering, University of Ljubljana,
Tržaška 25, SI-1000 Ljubljana, Slovenia
{rok.gajsek,france.mihelic}@fe.uni-lj.si
{vitos,bostjanv}@luks.fe.uni-lj.si
http://luks.fe.uni-lj.si/
†Faculty of Arts, University of Ljubljana
Aškerčeva 2, SI-1000 Ljubljana, Slovenia
{anja.podlesek,luka.komidar}@ff.uni-lj.si

**Abstract.** The paper deals with the recording and the evaluation of a multi modal (audio/video) database of spontaneous emotions. Firstly, motivation for this work is given and different recording strategies used are described. Special attention is given to the process of evaluating the emotional database. Different kappa statistics normally used in measuring the agreement between annotators are discussed. Following the problems of standard kappa coefficients, when used in emotional database assessment, a new time-weighted free-marginal kappa is presented. It differs from the other kappa statistics in that it weights each utterance's particular score of agreement based on the duration of the utterance. The new method is evaluated and the superiority over the standard kappa, when dealing with a database of spontaneous emotions, is demonstrated.

**Key words:** Emotion recognition, Kappa statistics, Emotional database

## 1 Introduction

Detecting emotions in speech has become a popular sub-field in speech recognition over recent years. Different tasks such as speech recognition, speaker identification or verification, development of dialog managers, etc. can benefit from added information about the psychological state of the speaker. A telephone based dialog system that would detect if a customer is getting annoyed and dissatisfied with the service could switch to a human operator for further assistance. In telecommunication systems a known emotional state of the person on the other side of the channel would enable a more credible exchange of information. These are just two possible use cases where emotion detection can be used.

A starting point for research in emotion recognition is a quality database. Most of available databases with emotional speech were obtained by recording professional actors as they portray particular emotion ([1, 2, 5, 4, 3]). Therefore they do not represent the situations in real life environment very adequately. Moreover to distinguish between normal (relaxed) and non-normal (emotional or aroused) conditions a fair amount of speech in the normal state is needed, which usually is not the case in emotional databases. Consequently, we decided to record our own audio and video database containing normal speech and spontaneous emotions [11].

In databases with acted emotions labelling of data is simplified since the actors are told which category of emotion to act out whereas with spontaneous emotions there is no a priori knowledge of the emotion expressed in the utterance. Different statistical measures can be used for assessing agreement between labelers, but non of them incorporates the duration of the utterance. We present a modified free-marginal multirater kappa [6] where duration of the utterance is taken into consideration.

The following paper firstly describes the planning and recording and then discusses the assessment of a database of spontaneous emotions. A new time-weighted kappa statistics focused on evaluation of emotional databases is also presented.

## 2 Recording strategies

Inducing spontaneous emotions in humans can only be achieved by some sort of deception. One option is to have a hidden camera type of experiment where a participant finds himself in a stressed situation. The advantage of such an approach is that the participant is not aware that he/she is being recorded and thus does not hold back his/her emotions as opposed to being in front of a camera. Drawbacks are that it is technically harder to carry out (audio and video recordings are usually lower quality), some participants do not want to give their consent, actors are needed to play out the scene, etc. For these reasons we decided to take a different approach. A scenario was constructed where participants would be told that the purpose of the experiment is to examine whether different biometric measures (audio, video, skin conductivity, heart beat) could be used in an adaptive test of intelligence. This sessions provided us with recordings of different levels of stress. For recording a particular emotion another scenario with videos and photographs targeting each of the main emotions was developed.

### 2.1 Recording session 1: Adaptive IQ test

Recording session called adaptive IQ test consisted of four parts. In the first part intended to record a normal state, a number of photographs with neutral content were presented to the participants. They were instructed to describe each photograph in detail, as if he/she would be describing it to a blind person. In this part we supposedly measured the verbal fluency.

In the second part the participants played a game of Tetris but instead of using the keyboard as input they lead the experimenter through the game uttering commands ”left”, ”right”, ”around” and ”down” (actual commands spoken were slovenian translations of these words: ”levo”, ”desno”, ”okrog” and ”dol”). The intent of this part was supposedly to assess the efficiency of his/her verbal instructions given to a teammate in order to achieve a common and specific goal.

Next, the game of Tetris that the participant played in the second part was played back and he/she had to describe what was happening on the screen by using the same four commands.

The final part of the session was the so-called adaptive intelligence test (adaptive meaning that the difficulty of the task will be chosen by the computer according to the correctness of the previous answers), the mental strategies used for solving the problem and the biometric measures. The test consisted of twenty three by three matrices where one element was missing and the participant had to determine which of the six offered answers logically fills the missing spot. The participants had to reason aloud about the principles of the arrangement of the matrix and the logic behind their answer supposedly for the psychologist to assess their mental strategy, but the real goal was to record speech under stress. Further pressure was put on the participants with additional information presented on the screen: current IQ value (constantly dropping regardless of the answers), heart beat (increasing throughout the experiment) and time left to finish the task. Analysis of subjective reports for 15 participants is presented in Fig 1.

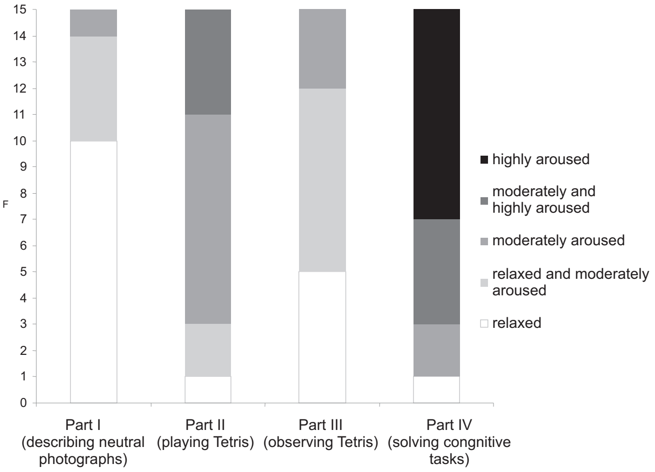

**Fig. 1.** Levels of arousal during different parts of the experiment based on the participants subjective report.

## 2.2 Recording session 2: Emotional videos and pictures

In the second session the participants watched a short video (approximately 10 minutes) and observed a set of photographs both targeting a particular emotion. After which they presented their thoughts on what they saw, how the videos made them feel, if something from the video or photographs relates to the situations in their life, etc. Emotions covered were happiness, anger, surprise, disgust, fear and sadness.

### 2.3 Database description

We recorded 19 participants in the adaptive IQ test session (12 women and 7 men) and 9 participants in the emotional videos session so far. Together that comprises approximately 30 hours of recorded material from which about a third is actual participants speech. The majority of the recordings have already been both transcribed and labelled according to the emotion.

## 3 Emotion labelling

With emotions acted out by professional actors there is usually not a big mismatch between the actor's interpretation and the annotator's opinion. This is expected since good acting means exactly that, adequately represent a particular emotion to the observer. But with spontaneous emotions, that appear at any time during a recording and last a random amount of time, the job of emotion labelling becomes more challenging. If a recording is split into shorter utterances (e.g. sentences), the annotator can focus on one at the time, but this way the information of the context in which the sentence was spoken is lost. Hence in our case the labelers were given a full recording and they were free to put the emotion labels anywhere and of any duration. The utterances defined by transcriptions were extracted from the recording and by comparing different annotator's scores the predominant label was chosen.

### 3.1 Measuring annotators agreement

The problem we are faced with when annotating emotional databases is that there are no references for different types of emotions. Therefore we presume each annotator is partly right. In order to combine their individual labels into a final one we have to examine the agreement between all of them. A predominant way of measuring the agreement between annotators in emotional databases is by using the kappa statistics [10].

### 3.2 Kappa coefficients

Different forms of kappa coefficients exist but are all based on the idea by Cohen [8] and shown in Eq. (1).

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

$P_e$ is an agreement between the annotators expected by chance and $P_o$ is the actual proportion of agreement between them. Thus the coefficient calculated from Eq. (1) represent the ratio between measured agreement and chance. If the value of $k$ equals zero this represents a level of agreement equal to chance, values above are better than chance and values below are worst. All various forms of kappa are based on this equation, but the differences between them lie in definitions of $P_o$ and $P_e$.

The most frequently used kappa coefficient in emotion labelling (as well as other fields) was presented by Fleiss [7], where he generalised Cohen's kappa from two raters to any number of raters. Fleiss defines $P_o$ as

$$P_o = \frac{1}{Nn(n-1)}((\sum_{i=1}^{N}\sum_{j=1}^{k} n_{ij}^2) - Nn), \tag{2}$$

where $n_{ij}$ is the number of annotators that assigned to case $i$ the class $j$, $n$ represents the total number of annotators, $N$ is the number of cases (utterances) and $k$ is the number of classes (emotional states). $P_e$ is defined as:

$$P_e = \sum_{j=1}^{k}(\frac{1}{Nn}\sum_{i=1}^{N} n_{ij})^2. \tag{3}$$

The probability of agreement by chance as defined in Eq. (3) presumes that all the classes ($k$) are represented equally across all cases. If there is a strong prevalence of one type of class over the other the Fleiss' coefficient drops regardless of the number of labels that are identical between the annotators. As it is demonstrated in [9] there is a quadratic relation between the value of Fleiss' coefficient and the prevalence of one type of class assuming other parameters are held constant ($n_{ij}$, $n$, $N$, $k$). When recording a database of spontaneous emotions one surely can not expect that all the different emotions or arousal states will be represented equally. Therefore the above calculations of Fleiss' kappa are not appropriate for measuring the agreement between annotators.

In [9] the author proposes a free-marginal multirater kappa as opposed to the standard Fleiss' kappa (which he calls Fleiss' fixed-marginal multirater kappa). The difference is that instead of using Eq. (3) for calculating the probability of agreement by chance he proposes that $P_e$ should be set constant to

$$P_e = 1/k, \tag{4}$$

where $k$ is the number of classes. This way there is no a priori restriction on the distribution of the classes and the above described problem of quadratic effect of prevalence of one class is lost (as it is graphically demonstrated in [9]). The equation of the new free-marginal multirater kappa coefficient thus becomes:

$$\kappa = \frac{(\frac{1}{Nn(n-1)}(\sum_{i=1}^{N}\sum_{j=1}^{k} n_{ij}^2 - Nn)) - \frac{1}{k}}{1-\frac{1}{k}} \tag{5}$$

Although we avoid the problem of distribution of the classes when using Eq. (5) there is a down side to this approach. The number of possible classes that can be used becomes important since the factor $1/k$ starts decreasing when the number of possible categories increases. Thus the coefficients also increase, regardless of the fact that the data does not change ($n_{ij}$, $n$, $N$).

In order to use the free-marginal multirater kappa in emotional labelling the number of different emotions available to the annotator needs to be carefully selected. Otherwise the values of the coefficient will be always close to 1, not reflecting the actual state. Also if the categories of emotions are equally presented in recordings (as is the case in most acted databases) the author in [9] suggests the use of standard Fleiss' kappa.

### 3.3 Time-wighted Free-Marginal Kappa

All kappa statistics described above handle each case equally. In speech databases these means that each utterance being labelled accounts equally to the final matching score regardless of the length. But in emotional databases the duration of the utterance that contains a particular emotion is important. This is the case with spontaneous emotions especially where the normal state usually prevails. Moreover if the annotators agree for example on a sentence lasting 15 seconds and disagree on a 2 second utterance, this means bigger agreement among them as if the case is the other way around. Therefore the duration of the utterance should be taken into consideration when measuring the agreement between annotators. We propose a modified kappa statistics focused specially on speech databases that incorporates the duration of the utterance. The equation for $P_e$ stays the same as for the free-marginal multirater kappa since emotional categories are not represented equally. The change is introduced to the calculation of $P_o$, where the original equation for one particular case (utterance)

$$P_i = \frac{1}{n(n-1)}(\sum_{j=1}^{k} n_{ij}^2 - n) \tag{6}$$

is not averaged over all the cases equally. Instead the average is calculated based on the duration of each case (utterance)

$$P_o = \frac{1}{T}\sum_{i=1}^{N} P_i t_i. \tag{7}$$

Equation for $P_o$ thus becomes:

$$P_o = \frac{1}{Tn(n-1)}\sum_{i=1}^{N}(\sum_{j=1}^{k} n_{ij}^2 - n)t_i, \tag{8}$$

where $T$ is the total time and $t_i$ is a duration of the utterance $i$.

## 4 Results

The new time-weighted free-marginal multirater kappa presented in the article was evaluated using a recording session from the AvID database. The new time-weighted free-marginal kappa is calculated in two cases: when emotions are classified in just two groups (normal/aroused) and the other when there are five different emotional categories. The values are compared against the standard Fleiss' kappa coefficients.

The quality of the different kappa coefficients cannot be judged by just comparing which score is higher. The superiority of one type needs to be evaluated theoretically and/or supported by some other measures. In order to provide some additional understanding of the annotators agreement, the Table 1 presents the percentage of time that the annotators agreed between each other. It shows that the annotators on average agree on the label around eighty percent of the time which seems like a good result, but the Fleiss' kappa coefficients presented in Table 2. give the impression that the scores are just above (and in some cases even below) chance.

| Annotator | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 5 | 78.50% | 76.97% | 84.01% | 84.08% |
| 4 | 79.19% | 78.19% | 76.53% | |
| 3 | 72.08% | 73.54% | | |
| 2 | 76.57% | | | |

**Table 1.** Agreement between five annotators in percentage of time.

Opposite to the standard Fleiss' Kappa coefficients the new time-weighted free-marginal Kappa values show a stronger agreement between the annotators. This correlates better with the time percentages from Table 1. The slight increase of the new time-weighted free-marginal kappa in the case of five emotional states comes from the above described effect of $1/k$ factor.

| Combinations of annotators | 1-2 | 1-3 | 1-4 | 1-5 | 2-3 | 2-4 | 2-5 | 3-4 | 3-5 | 4-5 | All 5 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Fleiss' Kappa | 0.35 | 0.09 | 0.34 | -0.02 | 0.21 | 0.32 | -0.04 | 0.12 | 0.04 | 0.05 | 0.17 |
| Time-weighted free-marginal Kappa (2 cat.) | 0.56 | 0.45 | 0.59 | 0.58 | 0.52 | 0.57 | 0.54 | 0.55 | 0.68 | 0.68 | 0.57 |
| Time-weighted free-marginal Kappa (5 cat.) | **0.72** | **0.66** | **0.75** | **0.74** | **0.70** | **0.73** | **0.71** | **0.72** | **0.80** | **0.80** | **0.72** |

**Table 2.** Kappa statistics between all combinations of two annotators and in the last column between all five.

## 5 Conclusion

In the paper we discussed the idea that the utterances should not be treated equally when evaluating agreement in speech databases. Instead the duration of the utterance should have an impact on the final score. Furthermore, Fleiss' kappa normally used for measuring agreement between the annotators, is not appropriate when working with databases of spontaneous emotions since all emotional categories are not represented equally. Therefore, a new time-weighted free-marginal kappa was introduced. Using our own emotional database AvID, discussed in the beginning of the paper, we evaluated the new kappa and compared it against the Fleiss' Kappa. The results show that the new time-weighted free-marginal kappa gives a more realistic measure of agreement between the annotators.

## References

1. LDC: SUSAS (Speech Under Simulated and Actual Stress). Language Data Consortium (1999) `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC99S78`
2. LDC: Emotional Prosody Speech and Transcripts. Language Data Consortium (2002) `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28`
3. Martin, O., Kotsia, I.,Macq, B., Pitas I.: The eNTERFACE05 Audio-Visual Emotion Database. Proc. of the 22nd International Conference on Data Engineering Workshops (ICDEW'06), pp. 8. (2006)
4. Burkhardt F.,Paeschke A., Rolfes M.,Sendlmeier W.,Weiss B.: A Database of German Emotional Speech. In Interspeech-2005, pp. 1517–1520. (2005)
5. Battocchi A., Pianesi F.: DAFEX: Un Database Di Espressioni Facciali Dinamiche. Proceedings of the SLI-GSCP Workshop "Comunicazione Parlata e Manifestazione delle Emozioni", (2004)
6. Randolph J. J.: Free-marginal multirater kappa: An alternative to Fleiss' fixed-marginal multirater kappa. Joensuu University Learning and Instruction Symposium 2005, Joensuu, Finland, (2005)
7. Fleiss J. L.: Measuring nominal scale agreement among many raters. Psychological Bulletin 76(5), pp. 378–382. (1971)
8. Cohen, J: A coefficient of agreement for nominal scales. Educational and Psychological Measurement, Vol. 20, No. 1, pp. 37–46. (1960)
9. Randolph, J. J.: Free-marginal multirater kappa: An alternative to Fleiss' fixed-marginal multirater kappa. Joensuu University Learning and Instruction Symposium 2005, Joensuu, Finland (2005)
10. Callejas Z., Lopez-Cozar R.: On the Use of Kappa Coefficients to Measure the Reliability of the Annotation of Non-acted Emotions. PIT '08: Proceedings of the 4th IEEE tutorial and research workshop on Perception and Interactive Technologies for Speech-Based Systems, pp. 221–232, Kloster Irsee, Germany (2008)
11. Gajšek R., Podlesek A., Komidar L., Socan G., Bajec B., Štruc, V., Bucik V., Mihelič F.: AvID : audio-video emotional database. Proceedings of the 11th International Multiconference Information Society - IS 2008, vol. C, pp. 70–74. Information Society, Ljubljana: Institute "Jožef Stefan" (2008)