CINEMO – A French Spoken Language Resource for Complex Emotions: Facts and Baselines

Björn Schuller, Riccardo Zaccarelli, Nicolas Rollet, Laurence Devillers

LIMSI-CNRS
Spoken Language Processing Group
BP 133, 91 403 Orsay cedex, France
(schuller|riccardo|nirollet|devil)@limsi.fr

Abstract
The CINEMO corpus of French emotional speech provides a richly annotated resource to help overcome the apparent lack of learning and testing speech material for complex, i.e. blended or mixed emotions. The protocol for its collection was dubbing selected emotional scenes from French movies. 51 speakers are contained and the total speech time amounts to 2 hours and 13 minutes and 4k speech chunks after segmentation. Extensive labelling was carried out in 16 categories for major and minor emotions and in 6 continuous dimensions. In this contribution we give insight into the corpus statistics focusing in particular on the topic of complex emotions, and provide benchmark recognition results obtained in exemplary large feature space evaluations. In the result the labelling off he collected speech clearly demonstrates that a complex handling of emotion seems needed. Further, the automatic recognition experiments provide evidence that the automatic recognition of blended emotions appears to be feasible.

1. Introduction

Emotion in real life is complex: we are not just ‘surprised’ or ‘angry’ or ‘joyful’, but pleasantly surprised or unpleasantly surprised. This is a self-evident and agreed upon fact (Devillers et al., 2005; Martin et al., 2006; Schröder et al., 2007) – yet, data for model construction and testing are sparse – at best (Campbell, 2003; Douglas-Cowie et al., 2003; Ververidis and Kotropoulos, 2003; Schuller et al., 2009b). And so are experiences with respect to obtain-able performances in automatic recognition of such blended emotions (Schuller et al., 2009b). Clearly, this is one of the next steps to be taken approaching machines’ human-like understanding of natural emotion following spontaneity and non-prototypicality (Schuller et al., 2009a).

The CINEMO corpus (Rollet et al., 2009) shall help to overcome this black hole in spoken language resources by provi-sion of labels for the ‘minor’, i.e. secondary, emotion in addi-tion to the ‘major’, i.e. primary or predominantly present, emotion together with the general mood. Moreover, next to such discrete categories it features a rich transcription with information in six dimensions consisting of activation and valence, appraisal-based control and suddenness, intensity, and naturalness. This annotation scheme has been proposed by the LIMSI according to manifold experience of annotation (Devillers et al., 2005; Schröder et al., 2007; Devillers and Martin, 2008), notably on appraisal dimensions (Devillers et al., 2006).

The CINEMO corpus contains acted emotional expression obtained by playing of dubbing exercises. This new proto-col (as described in detail in (Rollet et al., 2009)) is a way to collect mood-induced data (Gross and Levenson, 1995) in large amount which show several complex emotions.

The remainder of the paper is organized as follows: we first introduce the CINEMO corpus with according statistical figures in section 2. before detailing out the procedure and results of a baseline determination for complex emotions on the corpus in section 3. From these findings we draw conclusions and give future perspectives in section 4.

2. CINEMO Corpus Statistics
The CINEMO corpus (Rollet et al., 2009) features 3 992 instances after segmentation amounting to a total net play-time of 2:13:59 hours of emotional French speech by 51 speakers (21 female (1 656 instances), 30 male (2 336 instances)) in 4 age groups (–15 years, 15–25 years, 25–50 years, and 50+ years), of which none was a professional actor, captured by an on-board sound card and stored in 16 kHz, 16 Bit PCM to hard disk without conversion.

CINEMO’s general protocol is dubbing selected scenes that were picked from 12 French movies as depicted in Table 1 to encompass a broad coverage of emotions, provide situations close to everyday that feature the aimed at blend of emotions (Rottenberg et al., 2007), and are suited to sufficiently well induce mood (Gerrards-Hesse et al., 1994). Each of the overall 29 scenes could consist of one or two players at a time (14 male, 7 female, 6 mixed gender, 2 female–female scene and overall 119 turns ). By that overall all 31 roles are contained (14 female and 17 male).
The script of all scenes contains 119 turns with 1 609 words with 4.4 graphemes on average and a vocabulary size of 562 different terms. The distribution of N-grams is seen in table 2. The uni-grams “c” (this), “est” (is), and “j” (I) are the ones that appear more than 50 times, “c’est” is the bi-gram appearing more than ten times.

The participants had to superpose their voice on the actor’s one either with the latter audible or muted. In both cases the dialog as well as indications on pauses between the lines were shown on a screen as a Karaoke with the current word highlighted. The selected movie scenes are spoken interactions between two persons in everyday situations. We looked for natural contexts wherein the aimed at emotions

---

From: Proc. LREC 2010, ELRA, Valletta, Malta
| Film Title                      | # Year | Genre(s)                     |
|--------------------------------|--------|------------------------------|
| Astérix et Obélix: Mission Cléopâtre | 3 2002 | Adventure, Comedy, Family, Fantasy |
| Chaos                          | 2 2001 | Comedy, Drama, Crime         |
| Cité de la Peur                | 1 1994 | Comedy                       |
| Didier                         | 1 1997 | Comedy, Fantasy, Sport       |
| Escalier C                     | 6 1985 | Drama, Comedy                |
| Fauteuils d’orchestre          | 2 2006 | Comedy, Drama, Romance       |
| L’Auberge Espagnol             | 3 2002 | Comedy, Romance, Drama       |
| Le Corniaud                    | 2 1965 | Comedy, Crime                |
| Le goût des autres             | 2 2000 | Comedy, Drama, Romance       |
| Le héros de la famille         | 3 2006 | Drama                        |
| Le père Noël est une ordure   | 2 1982 | Comedy                       |
| Les tontons flingueurs         | 2 1963 | Comedy, Action, Crime        |

Table 1: Movies and number of scenes selected for dubbing.

| Frequency | # 1-gram | # 2-gram | # 3-gram | # 4-gram | # 5-gram |
|-----------|----------|----------|----------|----------|----------|
| ≥1        | 562      | 1331     | 1467     | 1422     | 1346     |
| ≥2        | 223      | 233      | 120      | 76       | 58       |
| ≥3        | 132      | 63       | 13       | 5        | 1        |
| ≥5        | 84       | 18       | 5        | –        | –        |
| ≥10       | 44       | 1        | –        | –        | –        |
| ≥50       | 3        | –        | –        | –        | –        |

Table 2: N-gram frequency in the CINEMO corpus 119 turns.

Each scene could be repeated, whereby the number of occurrences per attempt are 1 945 (first), 1 518 (second), 433 (third), 84 (fourth), 12 (fifth). By that each scene was repeated 1.67 times on average.

At present state it features a complete annotation by two experienced labelers ($L_1$: male, 31 years; $L_2$: female, 26 years). Two different strategies were intentionally followed: labeler $L_1$ was provided the context in sequential order and manually segmented the audio, whereas labeler $L_2$ was provided with single instances after segmentation in random order for verification. Segmentation was based on balancing interests between syntax, pragmatic, and stationarity of the major emotion, whereby shorter segments were preferred and predominant non-linguistic vocalizations served as additional segment-boundaries. The distribution of segment

(c.f. below) may occur (conjugal quarrel, receiving a present, feeling provoked, etc.). An example is:

- Movie, Scene: “Chaos”, cf. Figure 2.
- Type of affective state: sadness, disappointment (Role ‘A’ in Figure 2).
- Overall segment’s description: the speaker reports to her interlocutor the humiliating behavior of her boyfriend (who is not present in the scene)
- Involvement’s degree of the speaker: highly implicated, victim of conducts that are against commitments of fiancé’s
- Type of action or activity: storytelling
- Implied temporalities: recent past (few days)

Each scene could be repeated, whereby the number of occurrences per attempt are 1 945 (first), 1 518 (second), 433 (third), 84 (fourth), 12 (fifth). By that each scene was repeated 1.67 times on average.

At present state it features a complete annotation by two experienced labelers ($L_1$: male, 31 years; $L_2$: female, 26 years). Two different strategies were intentionally followed: labeler $L_1$ was provided the context in sequential order and manually segmented the audio, whereas labeler $L_2$ was provided with single instances after segmentation in random order for verification. Segmentation was based on balancing interests between syntax, pragmatic, and stationarity of the major emotion, whereby shorter segments were preferred and predominant non-linguistic vocalizations served as additional segment-boundaries. The distribution of segment

A: “Faut qu’je parle à quelqu’un, Fabrice me trompe .”
(I need to talk to someone, Fabrice is cheating on me.)
B: “Ah ?”
A: “Avec Charlotte une copine de fac .”
(With Charlotte - a friend from the faculty.)
B: “Ah zut !” (Oh no!)
A: “Et en plus elle est enceinte d’un autre mec .”
(And in addition, she is pregnant from another guy.)
B: “Qui ?” (Who?)
A: “Charlotte !”
B: “Ah !”
A: “Quand j’ai dit que j’as vu avec elle il a même pas nié .”
(When I told him that I’d seen him with her, he didn’t even deny.)
B: “Ah bon ?” (Really?)
A: “Non, il m’a dit bah ouais j’sors avec elle voilà .”
(No, he said to me, so yes, I go out with her, there you are.)
B: “Mince alors !” (Damn it!)
A: “Et pourtant on est fiancé officiellement .”
(And though we are officially engaged.)
B: “Bah oui .” (Oh yes.)
A: “Alors j’ai dit si c’est comme ça j’tes quitte et il m’a répondu fait comme tu penses .”
(Do as you think.)
B: “Ah !”
A: “Mais j’veux pas l’quitter !” (But I don’t want to quit him!)

Figure 1: Exemplary scene from the movie “Chaos”.
The following information is given per each of these instances: speaker ID and gender, movie ID, attempt (as described above), running ID and begin and end time for context preservation, as well as per labeler major and minor emotion attribute (16 options as shown in Table 3 with the numbers of instances per labeler), mood (7 options: amusement, irritation (ENN), neutrality (NEU), embarrassment (GEN), positivity (POS), stress, timidity (TIM)), whereby Cohen’s κ=0.41 for these labels (Cohen, 1968), as well as six three-state (i.e. low, middle, high or negative, neutral, positive) dimensions as detailed in Table 4 and in Figure 3 (selected). The imbalance in favor of negative valence is a typical observation in emotional databases. As the dimensions are ordinal, we also provide κ by linear and quadratic weighting next to Cohen’s unweighted κ for them. A monotonic increase going from unweighted to quadratic thereby indicates label confusions preferably in neighboring classes. Apart from suddenness, overall good degree of concurrence at κ ≥ 0.4 is observed. Note that we do not provide κ for the full 16 class per major and minor emotion paradigm: the reason is obvious – without reasonable clustering/grouping of labels or allowance of confusion between major and minor emotion (cf. (Devillers et al., 2006)) one can expect rather low values given this complexity. For transparency reasons we refrain from these steps herein.

Table 5 shows the distribution of minor emotions over major emotions as gray-scale heat map: each labeler’s frequency of minor labels per major label have been considered individually at first, and averaged afterwards. While of the potentially 256 class combinations only 118 are found in the set, the visualization clearly depicts the strong presence of blended emotions. If full agreement on major and minor emotion among the annotators is considered, 105 combinations remain with 2091 instances, i.e. half of the corpus – a further strong indication that the blended emotions are identifiable at a certain level of unambiguity.

### 3. Recognition of Complex Emotions

For acoustic modeling we use the openEAR toolkit’s (Eyben et al., 2009) “base” set of 988 features – a slight extension over the set provided for the INTERSPEECH 2009 Emotion Challenge (Schuller et al., 2009a) – which is extracted by systematic brute-forcing based on 19 functionals of 26 acoustic low-level descriptors (LLD, smoothed by simple moving average) and corresponding first order delta regression coefficients as depicted in Table 6. To further foster easy reproducibility of results and proper definition of a development set we decided for a straightforward three-fold partitioning by speaker index into train (≈40%/21 speakers: ID 1–21), development (≈30%/15 speakers: ID 22–36), and test (≈30%/15 speakers: ID 37–51). By that we ensure strict speaker independence and ‘genuine’ results w/o previous fine-tuning on the test partition.

To recognize complex emotions, we consider examples of major and minor emotions separately at first, which would resemble a maximum of 16 classes at a time in our case. In a second step, we directly target the complex compound in one classification pass. In theory this may lead to a quadratic number of classes, i.e. 256 in our case (cf. Ta-
Table 4: Number of instances per dimension from low (‘−1’) over middle (‘0’) to high (‘+1’) and labeler \((L_1,L_2)\) with according kappa values: unweighted \((\kappa_i, \text{i. e. Cohen’s})\), linear weighting \((\kappa^1)\), and quadratic weighting \((\kappa^2)\).

| # instances | \(-1\) | 0 | \(+1\) | \(kappa\) |
|-------------|-------|---|-------|---------|
| intenisty   | 1201  | 1234 | 2560  | 2574    | \(L_1\) | \(L_2\) |
| activation  | 667   | 727  | 2952  | 2870    | \(373\) | \(395\) |
| valence     | 2773  | 3115 | 351   | 249     | \(868\) | \(628\) |
| control     | 13    | 5    | 228   | 219     | \(3751\) | \(3768\) |
| suddenness  | 3907  | 3909 | 0     | 84      | \(83\)  | \(105\) |
| naturalness | 266   | 335  | 17    | 4       | \(3709\) | \(3653\) |

Table 5: Average distribution of minor emotions per major emotion for labelers \(L_1\) and \(L_2\). White to black resembles 0–100%.

| Minor | AMU | COL | DEC | ENE | INQ | IRO | JOI | NEG | NEU | POS | SAT | SED | STR | SUR | TRI |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| [%]   | 313 | 233 | 173 | 123 | 113 | 76  | 42  | 96  | 77  | 47  | 33  | 18  | 18  | 18  | 18  |

Table 6: Acoustic features: low-level descriptors and functionals. Abbreviations: Line Spectral Pairs (LSP), Mel Frequency Cepstral Coefficients (MFCC), Mean Absolute/Square Error (MAE/MSE).

| Example | WAR | UAR | AUC |
|---------|-----|-----|-----|
| ‘fixed minor’ | 48.1% | 50.7% | 0.710 |
| ‘fixed major’ | 53.2% | 44.6% | 0.711 |
| ‘fully mixed’ | 61.4% | 56.0% | 0.805 |

Table 7: Exemplary results by weighted and unweighted average recall (WAR, UAR) and area under the receiver operating curve (AUC) for three different five class constellations demonstrating CINEMO’s adequacy to research on complex emotions in increasing degree of blend and by that difficulty. Details in the text.
Table 8: Results for regression on selected dimensions based on mean of the labelers ($L_1$ and $L_2$).

| Example     | CC  | MLE |
|-------------|-----|-----|
| intensity   | 0.423 | 0.336 |
| activation  | 0.507 | 0.302 |

As last example we now choose a mixed task with overall 533 instances and accordingly 5 classes: INQ–NEU (114), STR–INQ (63), ENE–COL (186), ENE–DEC (110), and JOI–SUR (60). The observed result is again found in Table 7 (‘fully mixed’).

While these three examples are in no stricter relation to each other, they demonstrate that recognition of complex emotion seems feasible even in fully mixed presence of major and minor emotion not being fixed.

In Table 8 we provide additional results on regression for selected dimensions. The ground truth is obtained by the mean of the labelers (cf. Figure 3) and all instances are used.

Here we follow the popular metrics of cross correlation (CC) and mean linear error (MLE) (Grimm et al., 2007). To this end we shift to Support Vector Regression. Note that the prediction of these regressors can be used as features for the task of complex emotion recognition. Naturally, given the partly highly imbalanced distribution among the five discrete values that arise from the originally three given two labellers, performance is sub-optimal. A data-driven prequantization to have more balanced classes could change this in an elegant way but is not followed here, again for higher transparency.

4. Conclusion

In this work detailed statistics on the comparatively large (Schuller et al., 2009b) CINEMO corpus of complex emotions were given. In addition, we presented first impressions on the challenge of automatic recognition of the compound of emotions as encountered in everyday situations. An obvious direction for future research are tailored classification architectures that exploit the mutual information among major and minor emotions. In addition, complex ‘language models’ that not only reflect transition probability over time for a single, but for complex emotions, seem a promising next step. A self-evident precondition and desire in this respect are future large resources stemming from recordings ‘in the wild’.

5. Acknowledgement

This work was partly funded by the ANR project Affective Avatar.

6. References

N. Campbell. 2003. Databases of expressive speech. In Proc. Oriental COCOSDA Workshop.

J. Cohen. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 70(4):213–220.

L. Devillers and J.-C. Martin. 2008. Coding Emotional Events in Audiovisual Corpora. In Proc. LREC, Marrakech, Morocco.

L. Devillers, L. Vidrascu, and L. Lamel. 2005. Challenges in real-life emotion annotation and machine learning based detection. Neural Networks – Special Issue on “Emotion and Brain”, 18(4):407–422.

L. Devillers, R. Cowie, J.-C. Martin, E. Douglas-Cowie, S. Abrilian, and M. McRorie. 2006. Real-life emotions in French and English TV video corpus clips: an integrated annotation protocol combining continuous and discrete approaches. In Proc. LREC, Genoa, Italy.

E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach. 2003. Emotional speech: Towards a new generation of databases. Speech Communication, 40(1-2):33–60.

F. Eyben, M. Wöllmer, and B. Schuller. 2009. openEAR – Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In Proc. ACII. IEEE.

A. Gerrards-Hesse, K. Spies, and F.W. Hesse. 1994. Experimental inductions of emotional states and their effectiveness: A review. British Journal of Psychology, 85:55–78.

M. Grimm, K. Kroschel, E. Mower, and S. Narayanan. 2007. Primitives-based evaluation and estimation of emotions in speech. Speech Communication, 49(10-11):787–800.

J.J. Gross and R.W. Levenson. 1995. Emotion elicitation using films. Cognition & Emotion, 9(1):87–108.

J.-C. Martin, R. Niewiadomski, L. Devillers, S. Buisine, and C. Pelachaud. 2006. Multimodal Complex Emotions: Gesture Expressivity and Blended Facial Expressions. International Journal of Humanoid Robotics, 3(3):269–292.

N. Rollet, A. Delaborde, and L. Devillers. 2009. Protocol CINEMO: The use of fiction for collecting emotional data in naturalistic controlled oriented context. In Proc. ACII, Amsterdam, The Netherlands. IEEE.

J. Rottenberg, R.D. Ray, and J.J. Gross. 2007. Emotion elicitation using films. Series in Affective Science, pages 9–28. Oxford University Press.

M. Schröder, L. Devillers, K. Karpouzis, J.-C. Martin, C. Pelachaud, C. Peter, H. Pirker, B. Schuller, J. Tao, and I. Wilson. 2007. What Should a Generic Emotion Markup Language Be Able to Represent? In Ana Paiva, Rui Prada, and Rosalind W. Picard, editors, Affective Computing and Intelligent Interaction, pages 440–451, Berlin-Heidelberg. Springer.

B. Schuller, S. Steidl, and A. Batliner. 2009a. The INTER-SPEECH 2009 Emotion Challenge. In Proc. Interspeech, pages 312–315, Brighton, UK. ISCA.

B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth. 2009b. Acoustic emotion recognition: A benchmark comparison of performances. In Proc. Automatic Speech Recognition and Understanding Workshop (ASRU), Merano, Italy. IEEE. 13.-17.12.2009, to appear.

D. Ververidis and C. Kotropoulos. 2003. A review of emotional speech databases. In Proc. Panhellenic Conference on Informatics (PCI), pages 560–574, Thessaloniki, Greece.

I.H. Witten and E. Frank. 2005. Data mining: Practical machine learning tools and techniques, 2nd Edition. Morgan Kaufmann, San Francisco.