Incremental Acoustic Valence Recognition: an Inter-Corpus Perspective on Features, Matching, and Performance in a Gating Paradigm

Björn Schuller and Laurence Devillers

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

10.21437/Interspeech.2010-289

Abstract

It is not fully known how long it takes a human to reliably recognize emotion in speech from the beginning of a phrase. However, many technical applications demand for very quick system responses, e. g. to prepare different feedback alternatives before the end of a speaker turn in a dialog system. We therefore investigate this 'gating paradigm' employing two spoken language resources in a cross- and combined manner with a focus on valence: we determine how quick a reliable estimate is obtainable and whether matching by models trained on the same length of speech prevails. In addition we analyze how individual feature groups by type and derived functionals respond and find considerably different behavior. The language resources have been chosen to cover for manually segmented and automatically segmented speech at the same time. In the result one second of speech is sufficient on the datasets considered.

Index Terms: affective computing, automatic emotion recognition, incremental speech processing, gating paradigm

1. Introduction

Many real-time application scenarios of speech-based emotion recognition technology require instant estimates as soon as one starts talking. A good example are virtual agents, which need to prepare several responses that fit their human communication partner's emotion to trigger just the best one as soon as it is time to respond [1] – in fact like a human does when listening to his communication partner while already preparing different alternative responses to choose the best match when 'it is time'. This demands for incremental processing of speech, as has long been the case in the related field of Automatic Speech Recognition. In emotion recognition though, few works deal with the topic of different lengths and units, i. e. 'chunkings' of speech, and practically none systematically explores the reliability of a system output with increasing availability of speech – the 'gating paradigm' as known for word recognition (for an overview on this problematic cf. [2]). However, experiments on human perception considering facial (partly including speech) information exist, which demonstrate that already after 160 ms valence can be assessed after the start of its portray [3], whereby positive valence seems to be recognized earlier [4, 3]. In accordance to these named studies we also limit our following analyses to valence analysis, which at the same time seems very interesting from an application point of view. In music perception, it could also be shown that trained musicians are able to recognize a melody earlier [5], which is why we want to investigate not only how fast an emotion recognition system can sufficiently reliably detect emotion, but whether we can train it in this respect by matching the learning model to the available temporal context.

The remainder of the paper is organized as follows: we first introduce the French language resources used for our studies – the CINEMO and JEMO corpora – with according statistical figures on length distribution in section 2. The acoustic descriptors employed are detailed in section 3. Experiments and results are described in section 4. From these findings we draw conclusions and give future perspectives in section 5.

2. Affective Speech Corpora

In the following we introduce two different data sets that were selected to cover for manually segmented (CINEMO) and automatically segmented (JEMO) speech turns.

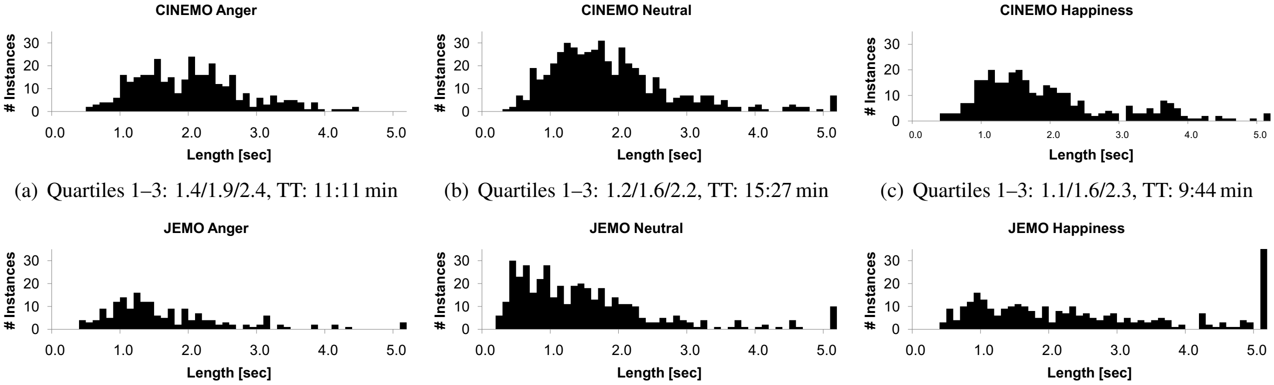
2.1. CINEMO

The CINEMO corpus [6] features 3 992 instances after segmentation amounting to a total net playtime of 2:13:59 h of emotional French speech by 51 speakers (21 female (1 656 instances), 30 male (2 336 instances)) in different age groups 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 to encompass a broad coverage of emotions in close to everyday situations, and induce mood sufficiently well [7]. The participants had to superpose their voice on the actor's 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. It features a complete annotation by two 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. Focusing on valence in this work, the following mapping was followed to select the ANGER, NEUTRAL, and HAPPINESS instances according to Table 1 from the whole CINEMO corpus which has a labeling for complex emotions by 16 'major' and 'minor' emotions and 6 dimensions (cf. [6] for details): all instances with an intensity rating of low and major emotion of neutral by both annotators were picked as NEUTRAL, all those with full

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

# Instances/Subjects		ANGER	NEUTRAL	HAPPINESS	sum	f	m	subjects f	subjects m	subjects f+m
	Train	230	407	166	803	354	449	20	15	35
CINEMO	**Test**	114	103	147	364	152	212	9	6	15
	Sum	344	510	313	1 167	506	661	29	21	50
	Train	119	284	222	625	301	324	14	13	27
JEMO	**Test**	60	132	94	286	144	142	7	5	12
	Sum	179	416	316	911	445	466	21	18	39

Table 1: *Number of instances and subjects per corpus, emotion, and gender.*



(a) Quartiles 1–3: 1.4/1.9/2.4, TT: 11:11 min

(b) Quartiles 1–3: 1.2/1.6/2.2, TT: 15:27 min

(c) Quartiles 1–3: 1.1/1.6/2.3, TT: 9:44 min

(d) Quartiles 1–3: 1.0/1.3/2.0, TT: 4:45 min

(e) Quartiles 1–3: 0.6/1.2/1.8, TT: 10:04 min

(f) Quartiles 1–3: 1.1/2.1/3.7, TT: 9:38 min

Figure 1: CINEMO and JEMO corpus: histogram of segment lengths per emotion. The rightmost bar cumulates all remaining (i. e. including all longer) instances (in the case of happiness this was cut off vertically in the image for the JEMO corpus. The full number resembles 52. Additionally quartiles 1–3 and the total time (TT) are given.

agreement on the major emotion belonging to satisfaction, joy, or amusement (whereby combinations of these were allowed) as HAPPINESS, and finally all those with cold or hot anger (again combinations allowed) as ANGER. For a better understanding of the experiments on duration from the beginning, the distributions of according segment lengths are shown in Figure 1.

2.2. JEMO

The JEMO corpus features speech recorded from 39 speakers (18 to 60 years old) disjunctive from those in CINEMO. The overall 1 134 instances amount to a total net playtime of 38:01 min. It was collected within an emotion-detection game where players had to act emotions such that the computer would recognize them. The game uses a speech segmentation based on silence pauses and an emotion recognizer based on the 5-emotions anger, fear, sadness, satisfaction and neutral built on CINEMO data. The linguistic content is free and the language French. The system detects the emotion plus activity (low or high) from the audio signal and depicts the detected emotion by an according emoticon on a screen visible to the player.

This game is a prototype of a real time detection system with recognition rate considerably below human performance (whereby anger and sadness are found on the upper and fear on the lower end). It is thus intended to motivate the players to depict the emotions. Such a paradigm might actually encourage the players to depict the emotion in a way that is consistent with what the recognizer is doing to classify it as when adjusting one's speaking style in order to be recognized by a speech recognizer. However, given the named below human performance this is rather unlikely. This recognizer performance often additionally led to a spontaneous and differentiated behavior (e. g. several negative reactions due to an emotion often not recognized and a positive reaction when the emotion was finally recognized were observed). Thus, speakers generated mixed acted and spontaneous utterances with higher level of expressivity than in CINEMO. Note that both these spontaneously occurring reactions, and the emoted acts are grouped together in the further analysis. The speech in the JEMO corpus has been annotated by the same two labelers as for CINEMO with respect to major and minor emotion independent of the actual game state. For our consideration of valence we focus on ANGER, NEUTRAL, and JOY based on the major emotion as for CINEMO, which mostly stem from the spontaneous player reactions. Distributions are provided accordingly in Table 1 and Figure 1.

3. Acoustic Features

For acoustic modeling we use the openSMILE toolkit's *"base"* set of 988 features – a slight extension over the set provided for the INTERSPEECH 2009 Emotion Challenge [8]. This set is extracted by systematic brute-forcing based on 19 functionals of 26 acoustic low-level descriptors (LLD, smoothed by simple moving average) well known to carry information on emotional state [9] and corresponding first order delta regression coefficients as depicted in Table 2 plus speaker gender as feature.

4. Experiments and Results

To foster easy reproducibility of results and proper definition of sets we decided for a straight-forward partitioning by speaker index into test (≈30 % / first (JEMO) and last (CINEMO) 12

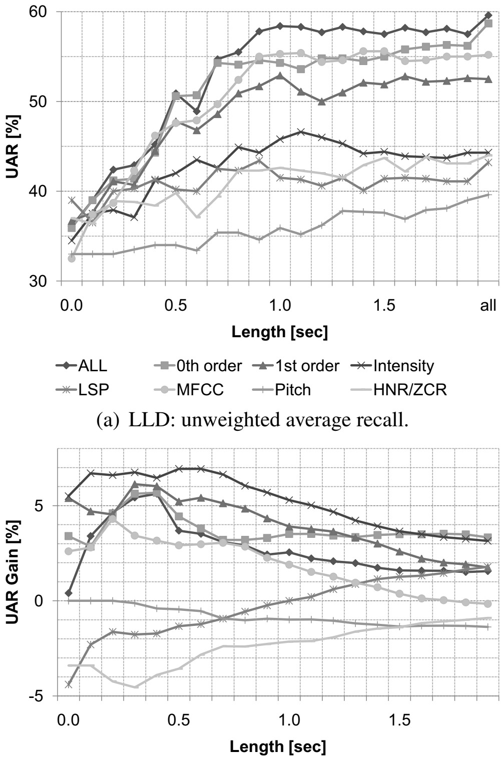
LLD (26 · 2)	**Functionals** (19)
(δ) Intensity	*Moments (4):*
(δ) Loudness	absolute mean, std. deviation
(δ) LSP Frequency 0–7	kurtosis, skewness
(δ) MFCC 1–12	*Extremes (3)/Positions (2):*
(δ) Pitch	2 × values, range/2 × position
(δ) Pitch envelope	*(Linear) Regression (4):*
(δ) HNR	offset, slope, MAE, MSE
(δ) ZCR	*Quartiles (6):*
	3 × quartiles, 3 × ranges

Table 2: *Acoustic features: low-level descriptors (LLD) and functionals. Abbreviations: Harmonics-to-Noise-Ratio (HNR), Line Spectral Pairs (LSP), Mel Frequency Cepstral Coefficients (MFCC), Mean Absolute/Square Error (MAE/MSE), Zero-Crossing-Rate (ZCR).*

speaker IDs per set), and train, whereby train is divided into two to optimise the classifiers, but united for the final experiments. By that we ensure strict speaker independence and 'genuine' results without previous fine-tuning on the test partition. The classifier of choice in this work are popular Support Vector Machines parametrized as polynomial Kernel with pairwise multi-class discrimination based on Sequential Minimal Optimization learning [10]. Parameters were optimized on the development sets and found optimal at 0.2 complexity and 1.0 exponential factor. Results are provided by unweighted average (UAR, to better reflect imbalance of instances among classes) or weighted average (WAR) recall (i. e. accuracy per class) as used in [8]. However, as classes are comparably balanced, no strong differences are observed in comparison to the weighted average recall (i. e. recognition rate).

4.1. Recall in Dependence of Gating

In our first experiment we consider the development of performance with increasing time on the combination of the two corpora to provide results more independent of manual or automatic segmentation (per train, development, and test partitions separately). We choose equidistant measure points in 0.1 s intervals for testing instances' time from the beginning, i. e. partial segments of up to a certain length as 0.1 s, 0.2 s, etc. were used to extract the features while the full segments are used to extract the features used in training as would usually be the case in a typical emotion recognizer. We consider lengths until 2.0 s and finally the whole chunk. Lengths in between are not followed, as no big differences are observed after 2.0 s. We always keep the number of testing instances fixed, meaning that chunks shorter than the current length of interest may be included. This is done per feature group (cf. Table 2), once 'horizontally' per low-level-descriptor group (intensity and loudness, pitch and its envelope, and HNR and ZCR are grouped together, the other groups are straight forward from the Table), once 'vertically' per functional type (as in italics in the table). In addition we consider all features together, and only 0^{th} order deltas (i. e. the actual features), and 1^{st} order deltas isolated. Figures 2a and 3a depict the according developments of unweighted average recall over time. As for the low-level descriptors, the figures do not only reveal the ranking of these, but different 'gradient' behavior: pitch is found among the flattest curves with increasing speech available, while e. g. the MFCC curve is comparably steep before the saturation at approximately 1 second. In the case of functionals only positions show a clearly flat curve. Note however that the curves also partly cross each other.



(a) LLD: unweighted average recall.

(b) LLD: absolute gain by matched length training.

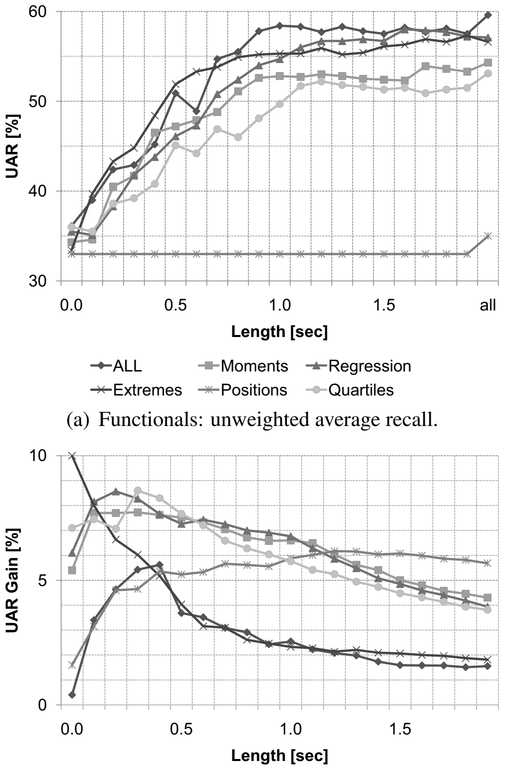
Figure 2: CINEMO and JEMO corpus combined.

4.2. Gain by Matched Length Training

We next consider how matching the lengths of the chunks in the learning model can improve the performance which is shown in Figures 2b and 3b. This means that we do not only gate the test, but accordingly the training instances. A real-life system can ensure such matched condition provided a reliable speech onset detection. Once onset is detected, it would need to 'switch' the learning models according to the current runtime of the speech chunk. In the figures the average UAR gain is shown over time in 'gating manner', i. e. the average at a time is calculated considering summing over previous measure points in 0.1 s steps.

Here, strong differences are observed for the low-level descriptors: while e. g. HNR/ZCR suffers from matching, Intensity clearly benefits from matching, in particular at the beginning of a speech chunk. Also in the case of functional-based analysis different behavior is observed, yet, grouped as functionals all benefit from matching. As examples, extremes show a monotonic average gain decrease pattern with evolving speech chunk length, while they benefit most from matching at the beginning. In contrast, positions in time of extremes show an almost monotonic respective increase pattern – naturally the curve has to fall at some point, as with increasing window length decreasing difference between matching and whole chunks is present.

We now investigate the effect of merged, intra- or inter- (i. e. cross-) corpus analysis looking also at automatic and manual segmentation separately. Table 3 first depicts results for WAR and UAR when using the whole chunks for training and at selected



(a) Functionals: unweighted average recall.

(b) Functionals: absolute gain by matched length training.

Figure 3: CINEMO and JEMO corpus combined.

lengths for testing. A remarkable gap in performance exists in intra-corpus investigation (JEMO is the apparent 'easier' task – as said potentially also owed to the subjects emoting in a way that addresses the feedback of the emotion recognition system, i. e. they are adjusting their expression to be consistent with what the recognizer is picking on); the merging of the training and testing partitions is found in between. Interestingly, shorter testing sequences can lead to better results for JEMO. One explanation might be labelers making their judgment rather early when segments are not 'pure' in their emotion. As to be expected, inter-corpus tests reveal considerably lower performance due to the different nature of the data-sets. The table next shows the average UAR gain by matching at selected four points in time as before. It is striking that only the CINEMO corpus benefits from matching, naturally in particular at the beginning. As JEMO apparently does not benefit from matching, the merged test produces a lowered gain over considering only CINEMO. This could be owed to the higher quality of manual segmentation.

5. Conclusion

In this work we investigated the accuracy increase starting from the beginning of a speech turn with increasingly available speech material. As a 'rule-of-thumb' number we observed one second of speech to be sufficient on the data used – afterwards considerably low further gain is observed having the full speech chunk available. We further investigated whether matching the learning model to the duration of the test size is beneficial. This was the case for one corpus (CINEMO), where segmentation was done manually. In a cross-corpus investigation length matching has proved to be counter-productive. Clear differences were further found for different feature and functional groups.

[%]	**Merge**	**Intra**		**Inter**	
Train/Test	CJ/CJ	C/C	J/J	J/C	C/J
WAR					
all	67.7	60.4	77.3	47.5	60.5
UAR					
≤0.5 s	55.7	48.0	68.2	39.1	56.2
≤1.0 s	66.0	57.5	72.6	43.5	57.8
≤1.5 s	66.3	57.8	74.4	45.3	59.0
≤2.0 s	65.5	60.6	72.9	46.3	56.5
all	67.0	60.9	72.0	47.8	56.4
UAR Gain by Matched Length Condition					
≤0.5 s	4.1	6.9	-0.1	-6.6	-8.8
≤1.0 s	1.3	2.5	-1.7	-4.1	-5.2
≤1.5 s	0.7	0.7	-0.6	-2.5	-4.2
≤2.0 s	0.2	-0.2	-1.1	-1.9	-3.4

Table 3: *Speaker independent recognition performance and average UAR gain for merged, intra-, and inter-corpus settings, all features. Abbreviations: CINEMO (C), JEMO (J).*

In a future investigation other quantizations can be considered, as voiced- or unvoiced segments instead of fixed length intervals and the behavior of gating for categorization of corpora.

6. References

[1] M. Schröder, R. Cowie, D. Heylen, M. Pantic, C. Pelachaud, and B. Schuller, "Towards responsive sensitive artificial listeners," in *Proc. 4th International Workshop on Human-Computer Conversation*, Bellagio, Italy, 2008.

[2] A. Batliner, D. Seppi, S. Steidl, and B. Schuller, "Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach," *Advances in Human Computer Interaction – Special Issue on "Emotion-Aware Natural Interaction"*, vol. 2010, Article ID 782802, pp. 15 pages, 2010.

[3] P. Barkhuysen, E. Krahmer, and M. Swerts, "Incremental perception of acted and real emotional speech," in *Proc. Interspeech*, Antwerp, Belgium, 2007, ISCA.

[4] J. Leppänen and J. K. Hietanen, "Positive facial expressions are recognized faster than negative facial expressions, but why?," *Psychological Research*, vol. 69, pp. 22–29, 2004.

[5] S. Dalla Bella, I. Peretz, and N. Aronoff, "Time course of melody recognition: A gating paradigm study," *Perception and Psychophysics*, vol. 65, no. 7, pp. 1019–1028, 2003.

[6] B. Schuller, R. Zaccarelli, N. Rollet, and L. Devillers, "CINEMO – A French Spoken Language Resource for Complex Emotions: Facts and Baselines," in *Proc. LREC 2010*, Valletta, Malta, 2010, pp. 1643–1647, ELRA.

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

[8] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," in *Proc. Interspeech*, Brighton, UK, 2009, pp. 312–315, ISCA.

[9] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, and N. Amir, "Whodunnit - searching for the most important feature types signalling emotion-related user states in speech," *Computer Speech and Language – Special Issue on "Affective Speech in real-life interactions"*, 2010.

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