

The Swedish PF-Star Multimodal Corpora

Jonas Beskow, Loredana Cerrato, Björn Granström, David House

Magnus Nordstrand, Gunilla Svanfeldt[†]

KTH, Speech Music and Hearing, 10044 Stockholm - Sweden

+46 8 79008965

{ beskow, loce, bjorn, davidh, magnusn, gunillas }@speech.kth.se

ABSTRACT

The aim of this paper is to present the multimodal speech corpora collected at KTH, in the framework of the European project PF-Star, and discuss some of the issues related to the analysis and implementation of human communicative and emotional visual correlates of speech in synthetic conversational agents. Two multimodal speech corpora have been collected by means of an opto-electronic system, which allows capturing the dynamics of emotional facial expressions with very high precision. The data has been evaluated through a classification test and the results show promising identification rates for the different acted emotions. These multimodal speech corpora will truly represent a valuable source to get more knowledge about how speech articulation and communicative gestures are affected by the expression of emotions.

Keywords

Multimodal corpora collection and analysis, visual correlates of emotional speech, facial animation.

INTRODUCTION

Analysis and synthesis of human-like gestures, in particular synchronisations of synthetic gestures with speech output, is achieving growing attention in the development of embodied conversational agents [9,5]. One of the greatest challenges is to implement believable, trustworthy, pleasant and human-like synthetic agents. This involves, amongst other aspects, having the agents display appropriate conversational behaviour and suitable visual correlates of expressive speech.

Analysis and visual synthesis of emotional expressions is one of the main areas of interest of the European project PF-Star [10]. The project aims at establishing future activities in the field of multisensorial and multilingual communication (Interface Technologies) by providing technological baselines, comparative evaluations, and assessment of prospects of core technologies, which future research and development efforts can build from.

One of the main activities of the first phase of the project has been the collection of audio-visual speech corpora and the definition of annotation format. These multimodal corpora are intended to provide materials for the analysis and modelling of human behaviour to be implemented in synthetic animated agents.

The animated synthetic talking heads that have been developed in our group are based on parameterised, deformable 3D facial models, controlled by rules within a text-to-speech framework. The rules generate the parameter tracks for the face from a representation of the text, taking coarticulation into account. A generalised parameterisation technique to adapt a static 3D-wireframe of a face for visual speech animation is applied [1]. This approach gives great freedom when it comes to making the synthetic faces expressive and having them perform gestures. However manual tailoring of facial gestures and emotional expressions can lead to unnaturalness of the synthesis and result in cartoon-like expressions. One way to avoid this is to obtain data that capture the dynamics of communicative and emotional facial expressions with very high precision. By capturing the facial movement of humans we can gain valuable insight into how to control the synthetic agent's facial gestures. To this end, multimodal speech corpora have been collected, and the aim of this paper is to present the different approaches of this acquisition as well as the content of the corpora and further discuss some of the issues related to the analysis and implementation of communicative and emotional visual correlates of human behaviour in synthetic conversational agents.

DATA COLLECTION

In order to be able to automatically extract relevant facial movements a motion capture procedure was employed. The data acquisition was carried out using an opto-electronic system - Qualysis MacReflex Motion Tracking System – [11] which allows capturing the dynamics of emotional facial expressions with very high precision.

Both articulatory data as well as other data related to facial movements can be recorded simultaneously, and the

[†] Authors in alphabetic order.

accuracy in the measurements is good enough for re-synthesis of an animated head (estimated mean error below 0.1 mm).

The data acquisition and processing is similar to earlier facial measurements carried out by [3,4]. Attaching infrared reflecting markers to the subject's face (as shown in figure 2) enables the system to register the 3D-coordinates for each marker at a frame-rate of 60Hz, i.e. every 17ms, by using four infrared cameras.

The utterances to be read and acted were displayed on a screen and recorded in one-minute chunks. Audio data was recorded on DAT-tape and visual data was recorded using a standard digital video camera and the optical motion tracking system.

Two corpora with two different non-professional actors have been collected with this set up:

- **corpus 1** consists of sample recordings aimed at evaluating the feasibility of different elicitation techniques such as reading prompts and interactive dialogue,
- **corpus 2** consists of non-sense words and short sentences, providing good phonetic coverage.

Corpus 1 consists of two sub-corpora, one of prompted speech and one of naturally elicited dialogues.

A total of 33 markers were used to record lip, eyebrow, cheek, chin and eyelid movements. Five markers attached to a pair of spectacles and three on the chest were used as a reference to be able to factor out head and torso movements.

The audio and visual data for the first sub-corpus was collected by having the speaker read prompted utterances, consisting of digit sequences and semantic neutral utterances, such as “*Linköping*” and “*ja*”.

Besides the seven universal prototypes for emotions: *anger*, *fear*, *surprise*, *sadness*, *disgust*, *happiness* and *neutral* [7], we asked the subject to act *worried*, *satisfied*, *insecure*, *confident*, *questioning*, *encouraging*, *doubtful* and *confirming*. These particular expressions were chosen since, in our opinion, they might be relevant in a future spoken dialogue system scenario. Some of these expressions were previously employed in the dialogue system Adapt [6].

The second sub-corpus consists of natural dialogues which were elicited using an information-seeking scenario. This communicative scenario is similar to one that might arise between a user and an embodied conversational agent in a dialogue system: there are two dialogue participants: A, who has the role of “information seeker” and B, who has the role of “information giver”. The domains of the dialogues were movie information (plots, schedules), and direction giving. The focus of the recording is on

participant B, the “information giver”, and only his movements were recorded (see Figure 1). However the audio recordings included the production of both subjects.



Figure 1 Data collection setup in corpus 1 with video and IR-cameras, microphone and a screen for prompts.

Corpus 2 consists of nonsense words and short sentences, providing good phonetic coverage. An actor was prompted with series of VCV, VCCV and CVCⁱⁱ nonsense words and short sentences, such as: “*grannen knackade på dörren*” (*the neighbour knocked on the door*). The sentences were kept content-neutral in order not to affect the acted expression. The actor was asked to produce them in six different emotional states, consisting of a sub-set of the expressions used in corpus 1, that is: *confident*, *confirming*, *questioning*, *insecure*, *happy*, and *neutral*. These particular expressions were selected since they are likely to be employed in dialogue systems. Some of these expressions can be interpreted pair wise on a positive-negative scale: confident versus insecure, confirming versus questioning. We did not include sad as opposed to happy and we did not include negative expressions such as anger, fear and disgust since they might not be appropriate expressions to be employed in a dialogue system.

A total of 35 markers were used to record lip, eyebrow, cheek, chin, and eyelid movements. Five markers attached to a pair of spectacles served as reference to factor out head moments (See figure 2).

Besides the natural dialogues in corpus 1, a total of 1700 items (i.e. words and sentences) were recorded. This material will provide the data for deriving statistically based models of the articulatory movements associated with expressive speech in Swedish. Part of this corpus has been used for a cross-evaluation test with the Italian partner of the PF-Star project. The test aims at comparing emotion recognition rates for Italian and Swedish natural (actor)

ⁱⁱ V= Vowel; C= Consonant

video sequences with those for Italian and Swedish synthetic faces [2].



Figure 2 Test subject in corpus 2, with IR-reflecting markers glued to the face.

DATA EVALUATION

A test was conducted to classify the data collected in corpus 2. A group of 13 volunteer Swedish students from KTH (6 female and 7 male) was presented with a total of 90 stimuli, consisting of digitised video-sequences of the Swedish actor uttering a random selection of the sentences in corpus 2 with the six expressions. The test was run in a plenary session, the stimuli were presented using a projected image on a wide screen, in random order, without the audio. Before the experimental session the participants were instructed to look at the video files and after each video-file select one of the seven options on the answering sheet, consisting of the six expressions and an extra category for “other”. The latter was inserted to avoid forced choice and a possible over-representation of neutral.

The percentages have been calculated on 78 stimuli, the first and last six stimuli responses were “dummies”.

The results are shown in the confusion matrix in Table 1, where the responses for other and no response have been collapsed in one column. On average 7% of each subject’s responses fell into these two categories.

All the expressions are identified above chance level, which means that the proportion of times that the subjects correctly identify the emotions is higher than the proportion of times one would expect identification by chance. No significant differences between the responses given by female and male subjects were found.

Happy and *neutral* (which are two of the basic emotions according to Ekman [7]) show much higher identification rates compared to the other expressions. *Confirming* gets 50% identification rate, and this is probably due to the fact that the actor typically produces head nods when acting this expression.

The main confusion seems to occur for *uncertain*, which

has been misidentified 41% of the times as *questioning*. However, *questioning* has been misjudged as *uncertain* only 8% of the times. In fact the misjudgements for *questioning* appear to be more evenly distributed across all

		judged						
		Expression	hap	conf	cer	neu	unc	que
acted	Happy	85%	2%	1%	1%	2%	8%	1%
	Confirming	12%	50%	12%	22%	1%	0%	4%
	Certain	1%	12%	37%	24%	3%	7%	16%
	Neutral	1%	3%	13%	70%	3%	2%	8%
	Uncertain	0%	3%	2%	2%	46%	41%	7%
	Questioning	7%	13%	15%	22%	8%	29%	7%
	Other							

the other classes.

Table 1 Confusion matrix for the identification test

The confusion between *uncertain* and *questioning* might be due to the fact that it is not easy to discriminate between them on the basis of the visual cues only. These two expressions are quite similar in their meaning (an unsure person might appear questioning at the same time) and the actor’s visual interpretation of these two expressions is similar: the typical gesture he uses is in both cases: eyebrow frowning.

Notwithstanding the confusions, and given the fact that the subjects judged the expressions on the basis of the visual cues only (i.e. without the support of the audio information), we believe that these results can be interpreted as an indication that the material collected in our corpus represent a reliable source for the analysis and measurement of different emotional facial expressions.

DATA TRANSCRIPTION AND ANNOTATION

All nonsense words and short sentences in the two corpora are provided with a phonetic transcription, which was automatically performed by an automatic aligner [12].

For the dialogues it is necessary to perform manual transcriptions and annotation. This can be done by using a dedicated annotation tool, such as ANVILⁱⁱⁱ. The annotation with ANVIL is performed on a freely definable multi-layered (tracks) annotation scheme, which can be *ad hoc* defined to label non-verbal communicative and expressive behaviour. An appropriate coding scheme was created to code the visual correlates of expressive speech and their specific function in the given context [4]. Some effort was spent in transcribing, annotating and analysing human behaviour in the recorded dialogues.

ⁱⁱⁱ <http://www.dfki.de/~kipp/anvil/>

Data annotation is necessary in order to couple the video-data to the 3D-data.

EXPLOITATION OF DATA

One of the goals of the analysis of the material in our multimodal corpora is to enable reproduction of trustworthy facial gestures – both emotional and other communicative gestures – in a talking face, to be used in dialogue systems. When trying to transfer the human knowledge in expressing facial gestures to a talking face, several crucial questions arise, such as what are the most appropriate and absolutely necessary expressions to implement? How can facial expressions be measured? How can we capture the complex interactions among articulatory gestures, the labial and facial visual cues related to the expression of communicative and emotional behaviour and the acoustic correlates of emotions, including prosodic features such as fundamental frequency parameters, voice quality and intonation?

Traditionally, visual and speech acoustic cues (both segmental and supra-segmental) conveying emotions have been studied separately. One of the main challenges of the PF-Star project is to understand how speech articulation and communicative gestures are coordinated.

One example is labial movements, which are controlled both by the phonetic-phonological constraints and the configurations required for the encoding of emotions. A preliminary analysis has been carried out to quantify the labial articulatory parameters modifications induced by the different emotions. The results of the investigation have shown how a number of articulatory and facial parameters for some Swedish vowels vary under the influence of expressive speech gestures [8]. Inspired by these results, we aim at building statistical models describing the interactions between articulation and emotional expression, and intend to apply these models to our talking heads.

DISCUSSION AND FUTURE WORK

The multimodal speech corpora described in this paper are very specific and even if their dimensions are not so extensive (only two actors and relative few items recorded), they can be valuable sources to get more knowledge about how speech articulation and communicative gestures are affected by the expression of emotions.

In order to extend our corpora, we are going to carry out further data collection with the opto-electronic system. The main focus of the next acquisition will be on dialogic speech. This will give better insight in how speech articulation, facial communicative gestures and emotional expressions interact with each other in a dialogic situation and in a more spontaneous speech style than reading of prompted speech.

Further analysis will be carried out to quantify articulatory parameter modifications induced by the different emotional expressions. Moreover we will study whether certain facial

emotional expressions are difficult to produce at the same time as certain communicative gestures and speech articulations. The knowledge acquired by analysing the data can be used to drive our 3D-agents, in terms of non-verbal and verbal emotional behaviour, leading, hopefully to innovative implementation in audiovisual synthesis.

Acknowledgements

Special thanks to Bertil Lyberg for making available the Qualisys Lab at Linköping University. The PF-Star project is funded by the European Commission, proposal number: IST2001 37599. This research was carried out at the Centre for Speech Technology, a competence centre at KTH, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organisations.

REFERENCES

1. Beskow J, 2003, *Talking heads - models and applications for Multimodal speech synthesis*. PhD thesis, TMH/KTH.
2. Beskow J, Cerrato L, Costantini E, Cosi P, Nordstrand M, Pianesi F, Prete M, Svanfeldt G, 2004, Preliminary Cross-cultural Evaluation of Expressiveness in Synthetic Faces, to appear in Proceedings of ADS 04.
3. Beskow J, Engwall O, Granström B, 2003, Re-synthesis of Facial and Intraoral Articulation from Simultaneous Measurements. *Proc. of ICPhS '03*. Barcelona, Spain, 57-60.
4. Cerrato L, Skhiri M, 2003, Analysis and measurement of communicative gestures in human dialogues, *Proc. of AVSP 2003*, St. Jorioz, France, 251-256.
5. DeCarlo D, Revilla C, Stone M, Venditti J, 2002 Making discourse visible: Coding and animating conversational facial displays *Computer Animation 02*, 11-16
6. Edlund J, Nordstrand M. 2002, Turn-taking Gestures and Hour-Glasses in a Multi-modal Dialogue System. *Proc. of ISCA Workshop Multi-Modal Dialogue in Mobile Environments*, Kloster Irsee, Germany, 181-184.
7. Ekman P, 1982, "Emotion in the human face" Cambridge University Press, New York.
8. Nordstrand M, Svanfeldt G, Granström B, and House D, (2003). Measurements of Articulatory Variation and Communicative Signals in Expressive Speech. *Proc. of AVSP'03*, 233-238.
9. Pelachaud C, Badler N, Steedman M., 1996, Generating Facial Expressions for Speech. *Cognitive Science 20*, 1-46.
10. PF-STAR: <http://pfstar.itc.it/> (Mars 04)
11. Qualisys: <http://www.qualisys.se> (Mars 04)

12.Sjölander K (2003). An HMM-based system for automatic segmentation and alignment of speech. *Proc. of Fonetik 2003* Umeå University, Department of Philosophy and Linguistics PHONUM 9, 93-96