Data-driven Generation of Emphatic Facial Displays

Mary Ellen Foster
Department of Informatics, Technical University of Munich
Boltzmannstraße 3, 85748 Garching, Germany
foster@in.tum.de

Jon Oberlander
Institute for Communicating and Collaborative Systems
School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, United Kingdom
jon@inf.ed.ac.uk

Abstract

We describe an implementation of data-driven selection of emphatic facial displays for an embodied conversational agent in a dialogue system. A corpus of sentences in the domain of the target dialogue system was recorded, and the facial displays used by the speaker were annotated. The data from those recordings was used in a range of models for generating facial displays, each model making use of a different amount of context or choosing displays differently within a context. The models were evaluated in two ways: by cross-validation against the corpus, and by asking users to rate the output. The predictions of the cross-validation study differed from the actual user ratings. While the cross-validation gave the highest scores to models making a majority choice within a context, the user study showed a significant preference for models that produced more variation. This preference was especially strong among the female subjects.

1 Introduction

It has long been documented that there are characteristic facial displays that accompany the emphasised parts of spoken utterances. For example, Ekman (1979) says that eyebrow raises "appear to coincide with primary vocal stress, or more simply with a word that is spoken more loudly." Correlations have also been found between prosodic features and events such as head nodding and the amplitude of mouth movements. When Krahmer and Swerts (2004) performed an empirical, cross-linguistic evaluation of the influence of brow movements on the perception of prosodic stress, they found that subjects preferred eyebrow movements to be correlated with the most prominent word in an utterance and that eyebrow movements boosted the perceived prominence of the word they were associated with.

While many facial displays have been shown to co-occur with prosodic accents, the converse is not true: in normal embodied speech, many pitch accents and other prosodic events are unaccompanied by any facial display, and when displays are used, the selection varies widely. Cassell and Thórisson (1999) demonstrated that "envelope" facial displays related to the process of conversation have a greater impact on successful interaction with an embodied conversational agent than do emotional displays. However, no description of face motion is sufficiently detailed that it can be used as the basis for selecting emphatic facial displays for an agent. This is therefore a task for which data-driven techniques are beneficial.

In this paper, we address the task of selecting emphatic facial displays for the talking head in the COMIC[1] multimodal dialogue system. In the basic COMIC process for generating multimodal output (Foster et al., 2005), facial displays are selected using simple rules based only on the pitch accents specified by the text generation system. In order to make a more sophisticated and naturalistic selection of facial displays, we recorded a single speaker reading a set of sentences drawn from the COMIC domain, and annotated the facial displays that he used and the contexts in which he used them. We then created models based on the data from this corpus and used them to choose the facial displays for the COMIC talking head.

[1] http://www.hcrc.ed.ac.uk/comic/

The rest of this paper is arranged as follows. First, in Section 2, we describe previous approaches to selecting non-verbal behaviour for embodied conversational agents. In Section 3, we then show how we collected and annotated a corpus of facial displays, and give some generalisations about the range of displays found in the corpus. After that, in Section 4, we outline how we implemented a range of models for selecting behaviours for the COMIC agent using the corpus data, using varying amounts of context and different selection strategies within a context. Next, we give the results of two evaluation studies comparing the quality of the output generated by the various models: a cross-validation study against the corpus (Section 5) and a direct user evaluation of the output (Section 6). In Section 7, we discuss the results of these two evaluations. Finally, in Section 8, we draw some conclusions from the current study and outline potential follow-up work.

2 Choosing Non-Verbal Behaviour for Embodied Conversational Agents

Embodied Conversational Agents (ECAs) are computer interfaces that are represented as human bodies, and that use their face and body in a human-like way in conversations with the user (Cassell et al., 2000). The main benefit of ECAs is that they allow users to interact with a computer in the most natural possible setting: face-to-face conversation. However, to realise this advantage fully, the agent must produce high-quality output, both verbal and non-verbal. A number of previous systems have based the choice of non-verbal behaviours for an ECA on the behaviours of humans in conversational situations. The implementations vary as to how directly they use the human data.

In some systems, motion specifications for the agent are created from scratch, using rules derived from studying human behaviour. For the REA agent (Cassell et al., 2001a), for example, gesturing behaviour was selected to perform particular communicative functions, using rules based on studies of typical North American non-verbal displays. Similarly, the Greta agent (de Carolis et al., 2002) selected its performative facial displays using hand-crafted rules to map from affective states to facial motions. Such implementations do not make direct use of any recorded human motions; this means that they generate average behaviours from a range of people, but it is difficult to adapt them to reproduce the behaviour of an individual.

In contrast, other ECA implementations have selected non-verbal behaviour based directly on motion-capture recordings of humans. Stone et al. (2004), for example, recorded an actor performing scripted output in the domain of the target system. They then segmented the recordings into coherent phrases and annotated them with the relevant semantic and pragmatic information, and combined the segments at run-time to produce complete performance specifications that were then played back on the agent. Cunningham et al. (2004) and Shimodaira et al. (2005) used similar techniques to base the appearance and motions of their talking heads directly on recordings of human faces. This technique is able to produce more naturalistic output than the more rule-based systems described above; however, capturing the motion requires specialised hardware, and the agent must be implemented in such a way that it can exactly reproduce the human motions.

A middle ground is to use a purely synthetic agent—one whose behaviour is controlled by high-level instructions, rather than based directly on human motions—but to create the instructions for that agent using the data from an annotated corpus of human behaviour. Like a motion-capture implementation, this technique can also produce increased naturalism in the output and also allows choices to be based on the motions of a single performer if necessary. However, annotating a video corpus can be less technically demanding than capturing and directly re-using real motions, especially when the corpus and the number of features under consideration are small. This approach has been taken, for example, by Cassell et al. (2001b) to choose posture shifts for REA, and by Kipp (2004) to select gestures for agents, and it is also the approach that we adopt here.

3 Recording and Annotation

The recording script for the data collection consisted of 444 sentences in the domain of the COMIC multimodal dialogue system; all of the sentences described one or more features of one or more bathroom-tile designs. The sentences were generated by the full COMIC output planner, and were selected to provide coverage of all of the syntactic patterns available to the system. In addition to the surface text, each sentence included all of the contextual information from the COMIC

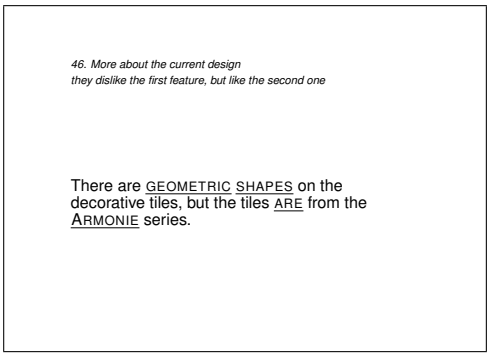


Figure 1: Sample prompt slide

planner: the predicted pitch accents—selected according to Steedman's (2000) theory of information structure and intonation—along with any information from the user model and dialogue history. The sentences were presented one at a time to the speaker, who was instructed to read each sentence out loud as expressively as possible while looking into a camera directed at his face. The segments for which the presentation planner specified pitch accents were highlighted, and any applicable user-model and dialogue-history information was included. Figure 1 shows a sample prompt slide.

The recorded videos were annotated by the first author, using a purpose-built tool that allowed any set of facial displays to be associated with any segment of the sentence. First, the video was split into clips corresponding to each sentence. After that, the facial displays in each clip were annotated. The following were the displays that were considered: eyebrow raising and lowering; eye squinting; head nodding (up, small down, large down); head leaning (left and right); and head turning (left and right). Figure 2 shows examples of two typical display combinations. Any combination of these facial displays could be associated with any of the relevant segments in the text. The relevant segments included all mentions of tile-design properties (e.g., colours, designers), modifiers such as *once again* and *also*, deictic determiners (*this*, *these*), and verbs in contrastive contexts (e.g., *are* in Figure 1). The annotation scheme treated all facial displays as batons rather than underliners (Ekman, 1979); that is, each display was associated with a single segment. If a facial display spanned a longer phrase in the speech, it was annotated as a series of identical batons on each of the segments.

Any predicted pitch accents and dialogue-history and user-model information from the COMIC presentation planner were also associated with each segment, as appropriate. We chose not to restrict our annotation to those segments with predicted pitch accents, because the speaker also made a large number of facial displays on segments with no predicted pitch accent; instead, we incorporated the predicted accent as an additional contextual factor. For the most part, the pitch accents used by the speaker followed the specifications on the slides. We did not explicitly consider the rhetorical or syntactic structure, as did, e.g., de Carolis et al. (2000); in general, the structure was fully determined by the context.

There were a total of 1993 relevant segments in the recorded sentences. Overall, the most frequent display combination was a small downward nod on its own, which occurred on just over 25% of the segments. The second largest class was no motion at all (20% of the segments), followed by downward nods (large and small) accompanied by brow raises. Further down the order, the various lateral motions appear; for this speaker, these were primarily turns to the right (Figure 2(a)) and leans to the left (Figure 2(b)).

The distribution of facial displays in specific contexts differed from the overall distribution. The biggest influence was the user-model evaluation: left leans, brow lowering, and eye squinting were all relatively more frequent on objects with negative user-model evaluations, while right turns and brow raises occurred more often in positive contexts. Other factors also had an influence: for example, nodding and brow raises were both more frequent on segments for which the COMIC planner specified a pitch accent. Foster (2006) gives a detailed analysis of these recordings.

4 Modelling the Corpus Data

We built a range of models using the data from the annotated corpus to select facial displays to accompany generated text. For each segment in the text, a model selected a display combination from among the displays used by the speaker in a similar context. All of the models used the corpus counts of displays associated with the segments directly, with no back-off or smoothing.

The models differed from one another in two ways: the amount of context that they used, and the way in which they made a selection within a context. There were three levels of context:

No context These models used the overall corpus counts for all segments.

(a) Right turn + brow raise (b) Left lean + brow lower

Figure 2: Typical speaker motions from the recording

Surface only These models used only the context provided by the word(s)—or, in some cases, a domain-specific semantic class. For example, a model would use the class DECORATION rather than the specific word *artwork*.

Full context In addition to the surface form, these models also used the pitch-accent specifications and contextual information supplied by the COMIC presentation planner. The contextual information was associated with the tile-design properties included in the sentence and indicated (a) whether that property had been mentioned before, (b) whether it was explicitly contrasted with a property of a previous design, and (c) the expected user evaluation of that property.

Within a context, there were two strategies for selecting a facial display:

Majority Choose the combination that occurred the largest number of times in the context.

Weighted Make a random choice from all combinations seen in the context, weighting the choice according to the relative frequency.

For example, in the no-context case, a majority-choice model would choose the small downward nod (the majority option) for every segment, while a weighted-choice model would choose a small downward nod with probability 0.25, no motion with probability 0.2, and the other displays with correspondingly decreasing probabilities.

These two factors produced a set of 6 models in total (3 context levels $\times$ 2 selection strategies). Throughout the rest of this paper, we will use two-character labels to refer to the models. The first character of each label indicates the amount of context that was used, while the second indicates the selection method within that context: for example, *SM* corresponds to a model that used the surface form only and made a majority choice.

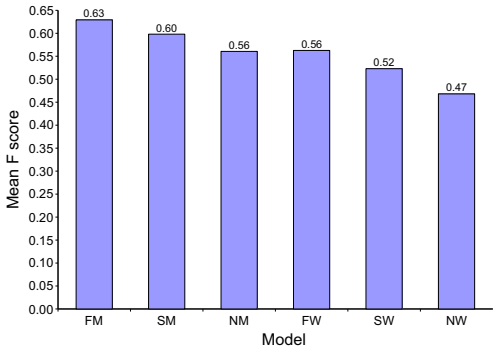


Figure 3: Mean F score for all models

5 Evaluation 1: Cross-validation

We first compared the performance of the models using 10-fold cross-validation against the corpus. For each fold, we built models using 90% of the sentences in the corpus, and then used those models to predict the facial displays for the sentences in the other 10% of the corpus. We measured the recall and precision on a sentence by comparing the predicted facial displays for each segment to the actual displays used by the speaker and averaging those scores across the sentence. We then used the recall and precision scores for a sentence to compute a sentence-level F score.

Averaged across all of the cross-validation folds, the *NM* model had the highest recall score, while the *FM* model scored highest for precision and F score. Figure 3 shows the average sentence-level F score for all of the models. All but one of the differences shown are significant at the $p <$

(a) Neutral (b) Right turn + brow raise (c) Left lean + brow lower

Figure 4: Synthesised version of motions from Figure 2

0.01 level on a paired T test; the performance of the *NM* and *FW* models was indistinguishable on F score, although the *FW* model scored higher on precision and the *NM* model on recall.

That the majority-choice models generally scored better on this measure than the weighted-choice models is not unexpected: a weighted-choice model is more likely to choose a less-common display, and if it chooses it in a context where the speaker did not, the score for that sentence is decreased. It is also not surprising that, within a selection strategy, the models that take into account more of the context did better than those that use less of it; this is simply an indication that there are patterns in the corpus, and that all of the contextual information contributes to the selection of displays.

6 Evaluation 2: User Ratings

The majority-choice models performed better on the cross-validation study than the weighted-choice ones did; however, this does not does not mean that users will necessarily like their output in practice. A large amount of the lateral motion and eyebrow movements occurs in the second- or third-largest class in a number of contexts, and is therefore less likely to be selected by a majority-choice model. If users like to see motion other than simple nodding, it might be that the schedules generated by the weighted-choice models are actually preferred. To address this question, we performed a user evaluation.

6.1 Experiment Design

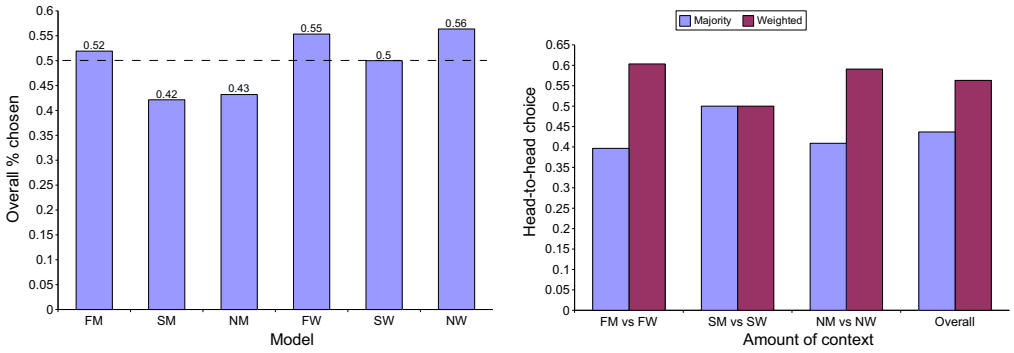
Materials For this study, we generated 30 new sentences from the COMIC system. The sentences were selected to ensure that they covered the full range of syntactic structures available to COMIC and that none of them was a duplicate of anything from the recording script. We then generated a facial schedule for each sentence using each of the six models. Note that, for some of the sentences, more than one model produced an identical sequence of facial displays, either because the majority choice in a broader context was the same as in a more narrow context, or because a weighted-choice model ended up selecting the majority option in every case. All such identical schedules were retained in the set of materials; in Section 6.2, we discuss their impact on the results. We then made videos of every schedule for every sentence, using the Festival speech synthesiser (Clark et al., 2004) and the RUTH talking head (DeCarlo et al., 2004). Figure 4 shows synthesised versions of the facial displays from Figure 2.

Procedure 33 subjects took part in the experiment: 17 female subjects and 16 males. They were primarily undergraduate students, between 20 and 24 years old, native speakers of English, with an intermediate amount of computer experience. Each subject in the study was shown videos of all 30 sentences in an individually-chosen random order. For each sentence, the subject saw two versions, each generated by a different model, and was asked to choose which version they liked better. The displayed versions were counterbalanced so that every subject performed each pairwise comparison of models twice, once in each order. The study was run over the web.

6.2 Results[2]

Figure 5(a) shows the overall preference rates for all of the models. For each model, the value shown

[2] We do not include those trials where both videos were identical; if these are included, the results are similar, but the distinctions described here just fail to reach significance.



(a) Overall preference rates

(b) Head-to-head preferences

Figure 5: User evaluation results

on that graph indicates the proportion of the time that model was chosen over any of the alternatives. For example, in all of the trials where the *FW* model was one of the options, it was chosen over the alternative 55% of the time. Note that the values on that graph should not be directly compared against one another; instead, each should be individually compared with 0.5 (the dotted line) to determine whether it was chosen more or less frequently than chance. A binomial test on these values indicates that both the *FW* and the *NW* models were chosen significantly above chance, while those generated by the *SM* and *NM* models were chosen significantly below chance (all $p < 0.05$). The choices on the *FM* and *SW* models were indistinguishable from chance.

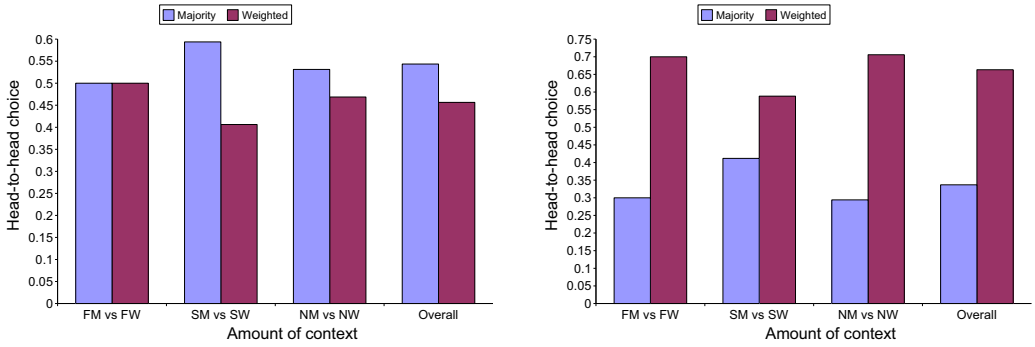
If we examine the preferences within a context, we also see a preference for the weighted-choice models. Figure 5(b) shows the preferences for selection strategy within each context. For example, when choosing between schedules both generated by models using the full context (*FM* vs. *FW*), subjects chose the one generated by the *FW* model 60% of the time. The trend in both the full-context and no-context cases is in favour of the weighted-choice models, and the combined values over all such trials (the rightmost pair of bars in the figure) show a significant preference for weighted choice over majority choice across all contexts (binomial test; $p < 0.05$).

Gender differences There was a large gender effect on the users' preferences: overall, the male subjects ($n = 16$) tended to choose the majority and weighted versions with almost equal probabilities, while the female subjects ($n = 17$) strongly preferred the weighted versions in any context, and chose the weighted versions significantly more often in head-to-head comparisons ($p < 0.001$). In fact, all of the overall preference for weighted-choice models came from the responses of the female subjects. The graphs in Figure 6 show the head-to-head preferences in all contexts for both groups of subjects.

7 Discussion

The predicted rankings from the cross-validation study differ from those in the human evaluation: while the cross-validation gave the highest scores to the majority-choice models, the human judges actually showed an overall preference for the weighted-choice models. This provides support for our hypothesis that humans would prefer generated output that reproduced more of the variation in the corpus, even if the choices made on specific sentences differ from those mode in the corpus. When Belz and Reiter (2006) performed a similar study comparing natural-language generation systems that used different text-planning strategies, they also found similar results: automated measures tended to favour majority-choice strategies, while human judges preferred those that made weighted choices. In general, this sort of automated measure will always tend to favour strategies that, on average, do not diverge far from what is found in the corpus, which indicates a drawback to using such measures alone to evaluate generation systems where variation is expected.

The current study also suggests a further drawback to corpus-based evaluation: users may vary systematically amongst themselves in what they prefer. All of the overall preference for weighted-choice models came from the female subjects;



(a) Male subjects

(b) Female subjects

Figure 6: Gender influence on head-to-head preferences

the male subjects did not express any significant preference either way, but had a mild preference for the majority-choice models. Previous studies on embodied conversational agents have exhibited gender effects that appear related this result: Robertson et al. (2004) found that, among schoolchildren, girls preferred a tutoring system that included an animated agent, while boys preferred one that did not; White et al. (2005) found that a more expressive talking head decreased male subjects' task performance when using the full COMIC system; while Bickmore and Cassell (2005) found that women trusted the REA agent more in embodied mode, while men trusted her more over the telephone. Taken together, these results imply that male users prefer and perform better using an embodied agent that is less expressive and that shows less variation in its motions, and may even prefer a system that does not have an agent at all. These results are independent of the gender of the agent: the COMIC agent is male, REA is female, while the gender of Robertson's agents was mixed. In any case, there is more general evidence that females have superior abilities in facial expression recognition (Hall, 1984).

8 Conclusions and Future Work

In this paper, we have demonstrated that there are patterns in the facial displays that this speaker used when giving different types of object descriptions in the COMIC system. The findings from the corpus analysis are compatible with previous findings on emphatic facial displays in general, and also provide a fine-grained analysis of the individual displays used by this speaker. Basing the recording scripts on the output of the presentation planner allowed full contextual information to be included in the annotated corpus; indeed, all of the contextual factors were found to influence the speaker's use of facial displays. We have also shown that a generation system that captures and reproduces the corpus patterns for a synthetic head can produce successful output. The results of the evaluation also demonstrate that female subjects are more receptive than male subjects to variation in facial displays; in combination with other related results, this indicates that expressive conversational agents are more likely to be successful with female users, regardless of the gender of the agent. Finally, we have shown the potential drawbacks of using a corpus to evaluate the output of a generation system.

There are three directions in which the work described here can be extended: improved corpus annotation, more sophisticated implementations, and further evaluations. First, the annotation on the corpus that was used here was done by a single annotator, in the context of a specific generation task. The findings from the corpus analysis generally agree with those of previous studies (e.g., the predicted pitch accent was correlated with nodding and eyebrow raises), and the corpus as it stands has proved useful for the task for which it was created. However, to get a more definitive picture of the patterns in the corpus, it should be re-annotated by multiple coders, and the inter-annotator agreement should be assessed. Possible extensions to the annotation scheme include timing information for the words and facial displays, and actual—as opposed to predicted—prosodic contours.

In the implementation described here, we built simple models based directly on the corpus counts and used them to select facial displays to accom-

pany previously-generated text; both of these aspects of the implementation can be extended in future. If we build more sophisticated *n*-gram-based models of the facial displays, using a full language-modelling toolkit, we can take into account contextual information from words other than those in a single segment, and back off smoothly through different amounts of context. Such models can also be integrated directly into the OpenCCG surface realiser (White, 2005)—which is already used as part of the COMIC output-generation process, and which uses *n*-grams to guide its search for a good realisation. This will allow the system to choose the text and facial displays in parallel rather than sequentially. Such an integrated implementation has a better chance at capturing the complex interactions between the two output channels.

Future evaluations should address several questions. First, we should gather users' opinions of the behaviours annotated in the corpus: it may be that subjects actually prefer the generated facial displays to the displays in the corpus, as was found by Belz and Reiter (2006). As well, further studies should look in more detail at the exact nature of the gender effect on user preferences, for instance by systematically varying the motion on different dimensions individually to see exactly which types of facial displays are liked and disliked by different demographic groups. Finally, if the extended *n*-gram-based model mentioned above is implemented, its performance should be measured and compared to that of the models described here, through both cross-validation and user studies.

Acknowledgements

Thanks to Matthew Stone, Michael White, and the anonymous EACL reviewers for their useful comments on previous versions of this paper.

References

A. Belz and E. Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *Proc. EACL 2006*.

T. Bickmore and J. Cassell. 2005. Social dialogue with embodied conversational agents. In J. van Kuppevelt, L. Dybkjær, and N. Bernsen, editors, *Advances in Natural, Multimodal Dialogue Systems*. Kluwer, New York.

B. de Carolis, V. Carofiglio, and C. Pelachaud. 2002. From discourse plans to believable behavior generation. In *Proc. INLG 2002*.

B. de Carolis, C. Pelachaud, and I. Poggi. 2000. Verbal and nonverbal discourse planning. In *Proc. AAMAS 2000 Workshop "Achieving Human-Like Behavior in Interactive Animated Agents"*.

J. Cassell, T. Bickmore, H. Vilhjálmsson, and H. Yan. 2001a. More than just a pretty face: Conversational protocols and the affordances of embodiment. *Knowledge-Based Systems*, 14(1–2):55–64.

J. Cassell, Y. Nakano, T. W. Bickmore, C. L. Sidner, and C. Rich. 2001b. Non-verbal cues for discourse structure. In *Proc. ACL 2001*.

J. Cassell, J. Sullivan, S. Prevost, and E. Churchill. 2000. *Embodied Conversational Agents*. MIT Press.

J. Cassell and K. R. Thórisson. 1999. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 13(4–5):519–538.

R. A. J. Clark, K. Richmond, and S. King. 2004. Festival 2 – build your own general purpose unit selection speech synthesiser. In *Proc. 5th ISCA Workshop on Speech Synthesis*.

D. W. Cunningham, M. Kleiner, H. H. Bülthoff, and C. Wallraven. 2004. The components of conversational facial expressions. In *Proc. APGV 2004*, pages 143–150.

D. DeCarlo, M. Stone, C. Revilla, and J. Venditti. 2004. Specifying and animating facial signals for discourse in embodied conversational agents. *Computer Animation and Virtual Worlds*, 15(1):27–38.

P. Ekman. 1979. About brows: Emotional and conversational signals. In M. von Cranach, K. Foppa, W. Lepenies, and D. Ploog, editors, *Human Ethology: Claims and limits of a new discipline*. Cambridge University Press.

M. E. Foster. 2006. *Non-default choice in generation systems*. Ph.D. thesis, School of Informatics, University of Edinburgh. In preparation.

M. E. Foster, M. White, A. Setzer, and R. Catizone. 2005. Multimodal generation in the COMIC dialogue system. In *Proc. ACL 2005 Demo Session*.

J. A. Hall. 1984. *Nonverbal sex differences: Communication accuracy and expressive style*. The Johns Hopkins University Press.

M. Kipp. 2004. *Gesture Generation by Imitation - From Human Behavior to Computer Character Animation*. Dissertation.com.

E. Krahmer and M. Swerts. 2004. More about brows: A cross-linguistic study via analysis-by-synthesis. In C. Pelachaud and Z. Ruttkay, editors, *From Brows to Trust: Evaluating Embodied Conversational Agents*, pages 191–216. Kluwer.

J. Robertson, B. Cross, H. Macleod, and P. Wiemer-Hastings. 2004. Children's interactions with animated agents in an intelligent tutoring system. *International Journal of Artificial Intelligence in Education*, 14:335–357.

H. Shimodaira, K. Uematsu, S. Kawamoto, G. Hofer, and M. Nakai. 2005. Analysis and synthesis of head motion for lifelike conversational agents. In *Proc. MLMI 2005*.

M. Steedman. 2000. Information structure and the syntax-phonology interface. *Linguistic Inquiry*, 31(4):649–689.

M. Stone, D. DeCarlo, I. Oh, C. Rodriguez, A. Lees, A. Stere, and C. Bregler. 2004. Speaking with hands: Creating animated conversational characters from recordings of human performance. *ACM Transactions on Graphics (TOG)*, 23(3):506–513.

M. White. 2005. Efficient realization of coordinate structures in Combinatory Categorial Grammar. *Research on Language and Computation*. To appear.

M. White, M. E. Foster, J. Oberlander, and A. Brown. 2005. Using facial feedback to enhance turn-taking in a multimodal dialogue system. In *Proc. HCI International 2005*.