Language-driven nonverbal communication in a bilingual conversational agent

Scott A. King, Alistair Knott and Brendan McCane
Dept of Computer Science
University of Otago
PO Box 56
Dunedin
New Zealand
+64 3 479 8579 Fax +64 3 427 8529
{sking, alik, mccane}@cs.otago.ac.nz

Abstract

This paper describes an animated conversational agent called Kare[1] *which integrates a talking head interface with a linguistically motivated human-machine dialogue system. The agent has a range of nonverbal behaviors, which involve a mixture of machine vision, computer animation and natural language processing techniques. The system's architecture couples the agent's nonverbal communicative processes very tightly to its model of verbal interaction. We discuss several consequences of this architecture, in particular the ability to use different non-verbal dialogue management signals when speaking different languages.*

1 Dialogue Management for Animated Conversational Agents

Over the last few years, computational linguists have become interested in using animated conversational agents as an interface medium with the user. Some of this interest centers around lip synchronization in speech synthesis [17, 6, 13]. Other researchers have developed agents which use nonverbal methods to realize aspects of the information structure and semantics of sentences [4, 5, 9]. Finally, a large number of researchers are interested in developing agents which participate in dialogues. The theoretical frameworks which are developed for these agents are based around models of face-to-face interaction, and focus on the nonverbal expression of turn-taking signals, signals accompanying dialogue acts and signals helping to convey propositional information [3], models of deixis [15] and of gesture [1], combining facial expressions of differing functions [18], and emotional expression and concealment [7].

In this paper, we describe how a dialogue management system originally designed purely for written text was extended to control the behavior of an animated conversational agent. The dialogue system is called Te Kaitito[2] [14, 8]: it supports conversation with the user in either English or Māori, in simple knowledge-authoring and information-seeking dialogues. The animated agent is called TalkingHead [13]: it is designed specifically to produce speech-synchronized animation, and it is capable of animating multiple characters using multiple languages.

Our project to link these two systems has highlighted two main points. Firstly, we are interested to what extent the model of discourse and dialogue developed for the purely linguistic application would suffice to generate the animated agent's nonverbal behavior. This issue is discussed in Section 2. Secondly, Te Kaitito can converse in two different languages: speakers of English and Māori use different nonverbal conventions, and the animated agent must be able to reproduce these differences. These differences are discussed in Section 3. Section 4 describes our implementation with some results presented in Section 5.

[1] Pronounced as in French *carré*. *Te Karetao* is Māori for 'puppet'. The shortened *Kare* is also a term of endearment.

[2] *Te Kaitito* is Māori for 'the composer', or 'the improviser'.

2 Architecture for the Conversational Agent

Te Kaitito is a collection of natural language processing (NLP) resources for English and Māori. The system is designed to include a module for all of the major tasks involved in the interpretation and generation of linguistic utterances, including sentence parsing and disambiguation, anaphora and presupposition resolution, dialogue management, and the planning and generation of single- or multiple-sentence responses. For our animated agent, we envisage an architecture in which Te Kaitito passes the talking head all the relevant verbal information it needs at key points in this processing, both during interpretation of the user's utterance, during dialogue management, and during response generation. We are interested to know what information the talking head might need in addition to these messages from Te Kaitito.

There are three kinds of information relevant to nonlinguistic signals computed by Te Kaitito:

- **Incoming dialogue act**. When the user gives Kare an utterance to process, Kare has to establish what **dialogue act** the user is executing. For example, from the conversation in Figure 1, Kare recognizes the following incoming dialogue acts: ASSERT (lines 1,3,11,13,15,17, and 19), YES/NO-QUESTION (lines 5, 7, and 9), and WH-QUESTION (lines 21, 23, and 25). The system also recognizes errors and generates appropriate responses in lines 4 and 6. The recognition of a particular dialogue act, or the registering of a particular interpretation error, are likely to have reflexes in nonverbal behavior.

- **Context updates** resulting from an incoming dialogue act. For instance, when a QUESTION is received, the system's representation of the dialogue context is updated with a new goal to answer this question. Again, these updates could have reflexes in nonverbal behavior.

- **Response structure**. This is the message with which the system responds to the user. It is more than just a piece of text; it is a linguistic object—essentially a piece of syntactic structure. In Te Kaitito, the structure contains the **response dialogue act** this utterance constitutes , the **language** the utterance is in, and its **information structure** (which parts are given and which bits are new, roughly speaking). For the conversation in Figure 1, the response dialogue act is one of the following: ACKNOWLEDGE (lines 2, 12, 14, 16, and 20), YES-ANSWER (line 10), NO-ANSWER (line 8), WH-ANSWER (lines 24 and 26), and CLARIFICATION-QUESTION (lines 18, 20, and 22). The response dialogue act will clearly be important for the nonverbal signals which accompany the speech. Information structure is important to specify the prosody and the associated nonverbal signals of the synthesized speech.

1 **User**: A big dog chased a black cat
2 **Kare**: Okay. - āna
3 **User**: The pimple barked
4 **Kare**: That makes no sense to me
5 **User**: Did the aardvark bark
6 **Kare**: I don't know. - kaore au i te paku mohio
7 **User**: Did the cat chase the dog
8 **Kare**: No - kao
9 **User**: Did the dog chase the cat
10 **Kare**: Yes - ae
11 **User**: A black dog was in a garden
12 **Kare**: Okay. - āna
13 **User**: The dog barked
14 **Kare**: Okay. - āna
15 **User**: A black dog was in a garden
16 **Kare**: Okay. - āna
17 **User**: The dog ate
18 **Kare**: Which one? - ko tehea
19 **User**: The big dog ate
20 **Kare**: Okay. - āna
21 **User**: What did the dog chase
22 **Kare**: Which one? - ko tehea
23 **User**: What did the big dog chase
24 **Kare**: It was the cat - na te ngeru
25 **User**: What chased the cat
26 **Kare**: It was the big dog - na te kuri nui

Figure 1. An example conversation with Kare. The responses from Kare are given here in both English and Māori. However, during a conversation the system responds in one language at a time, but that language can be changed during the conversation.

What control does the animated agent need apart from these sources of information? Certainly there are inputs which would be required if the agent was operating in an environment in which tasks other than face-to-face communication were performed (the kind of environments that STEVE [19] and Rea [3] operate in). But we are thinking about purely communicative, nonverbal operations. We believe that the linguistic information Te Kaitito already generates, as just outlined, comprises most of the information the talking head needs.

However, there are additional low-level channels of face-to-face interaction which we believe run on a completely different loop: for instance, postural congruence [20], or congruence of facial expression. Another plausible independent channel is one whereby an agent signals to the other that (s)he is still actively involved in the conversation. This involves orienting roughly towards the interlocutor. In other words, the talking head needs to keep track of the user's position. Note that the operation of this 'user-finding' system does not mean that the head has to be *gazing* at the user at all times; this is precisely one of the things which will be under the control of the verbal system.

3 Culture-specific Dialogue Conventions

There are some very clear differences in nonverbal communication conventions between English and Māori (and other Polynesian languages for that matter). These have been extensively documented anecdotally, and are well known as the source of cross-cultural communication difficulties. In a wide-ranging survey, Metge and Kinloch [16] describe several differences in non-verbal dialogue cues. We will discuss three such differences.

3.1 Nonverbal Signals for Agreement and Disagreement

Firstly, Polynesian speakers employ some distinctive signals for agreement, disagreement and acknowledgment. "[Polynesians] recognise the nod and headshake as *yes* and *no*, but commonly use other indicators: an upward movement of the head and/or eyebrows for *yes* and an unresponsive stare—straight ahead or down at the feet—for *no*. These are easily misread [by European New Zealanders]." [16]

The eyebrow flash for *yes*, or for acknowledgment dialogue acts, is indeed frequently misread. Eibl-Eibesfeldt [11] and Grammer *et al.* [12] confirm that this nonverbal signal has a very wide range of discourse and interpersonal meanings across cultures throughout the world.

3.2 Verbal/Nonverbal Overloading

It is sometimes possible to convey a message both verbally and nonverbally. For instance, to answer *yes* in English, a speaker can either nod, or say *yes*, or **overload**, by doing both. However, the choice as to which medium to use is also subject to cultural differences. "[European New Zealanders] usually say *yes* and *no*, reinforcing the words with a nod or a shake of the head. They accept the words without the action, but regard the actions without the words as inadequate and rude except in situations of intimacy. Maori and Samoans on the other hand frequently dispense with the verbal forms and rely on gestures only without considering this rude. " [16].

3.3 Eye Contact for Managing Dialogue

For American and British English the patterns of speaker and hearer gaze in dialogue are well known [10]. When the speaker is talking, (s)he looks at the hearer intermittently; when (s)he wishes to cede the conversational floor, (s)he gazes at the hearer more consistently. The listener gazes more at the speaker, especially when (s)he wishes to gain the floor. However, "Maori and Samoans consider it (...) impolite to look directly at others when talking to them. They say that it tends to put the two concerned into a relationship of conflict and confrontation. (...) So they rest their gaze elsewhere, slightly to one side, on the floor, ceiling or distant horizon, or they even close their eyes altogether." [16].

3.4 A Function for Nonverbal Signals

From the above observations, it makes sense to think of the appropriate nonverbal signals for an agent to generate as a function of (at least) the language being used and the dialogue act being performed. The following table describes a simple function approximating Metge and Kinloch's observations, and demonstrating the dependence of the agent's language of interaction on nonverbal signals.

Dialog act	Lang.	Action
Yes	English Māori	Nod. Eyebrow flash.
No	English Māori	Shake head. Shake head/look at feet.
Speaking	English Māori	Make eye contact. Avoid eye contact.
Accept assertion	English Māori	Nod and/or 'okay'. Eyebrow flash or 'āna'.

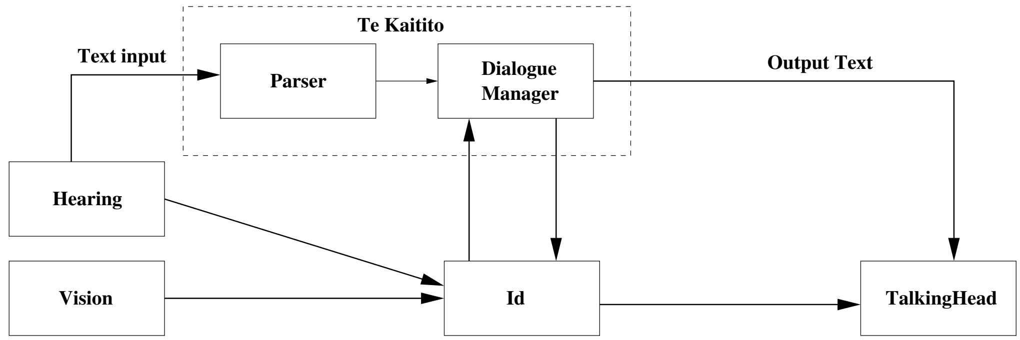


Figure 2. Overview of Kare.

4 Kare Overview

Kare is our implementation of a conversational agent for human computer interaction using the architecture of Figure 2.

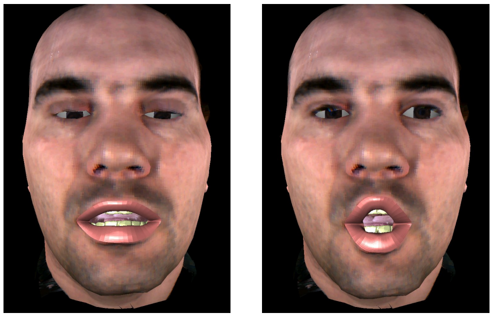
The system reacts to spoken discourse using standard speech recognition techniques. Currently, we use CMU Sphinx2 [21] in our system and we only recognize English. However, Māori and English can be input via the keyboard. The speech is converted to text and sent to Te Kaitito for processing. Te Kaitito first determines the type of dialogue act (question, assertion, acknowledgment, ...) and informs the Id module. The Id interfaces the various parts of the systems together, and gives Kare it's personality. The Id sends any appropriate response to TalkingHead, such as furrowing the brows and looking off in space if a question is asked. This is done for both Māori and English. Although there is a cultural reason to pause to collect one's thoughts in Māori, and the gesture of looking away may indicate to the listener that the speaker is concentrating on finding the answer, here we use the gesture to hide the delay in the system for processing.

The Id and Te Kaitito exchange information that will guide Te Kaitito in generating a response. It will also eventually use its personality to help Te Kaitito choose between possible responses. Te Kaitito then produces an appropriate response, for instance the answer to a posed question. The response is in the form of marked up text that is sent to TalkingHead for rendering. Note that the text may contain only nonverbal communication.

The Id controls the agent at a low-level performing tasks such as blinking and eye gaze. Between conversation acts these actions are performed by the Id without consulting Te Kaitito, and their purpose is to give life to the agent. During conversation acts, however, these actions may be overridden or synchronized with nonverbal gestures or speech. For instance, blinks occur automatically to keep the eye moist, but can be controlled consciously when staring intently to show interest in the speaker's words or synchronized with the beginning of words during speech. Sometimes eye gaze is controlled directly by the dialogue manager, for instance when forcing eye contact or avoiding eye contact. At other times, the Id controls the eyes directly, such as when the agent shakes its head the eyes may remain focused on a spot during the head shake.

The Id uses vision techniques to determine the location of the head of the interlocutor/human to control eye gaze. We use a consumer-grade webcam to take an image from the computer's viewpoint and we use the method of Viola and Jones [22] to locate faces in that image. This involves training a cascade of AdaBoost classifiers from a set of positive and negative images. The technique is appealing because it runs in real-time on standard PC hardware, and works well in an uncontrolled environment. The performance of the face detector has been promising, and initial experiments indicate that we can achieve a false positive rate of between 0.4% to 0.1% while maintaining a detection rate of greater than 95%. The false positive rate is still too high for excellent performance, but it should be adequate for our application under the right conditions.

TalkingHead [13] is a multi-lingual text-to-audiovisual-speech system that we use to embody Kare . TalkingHead takes the text from the Dialogue Manager and produces lip-synchronized animation. The audio is produced using Festival [2], a freely available, general, multi-lingual speech synthesis system. Facial expressions are generated from markup tags in the input text (such as (nod),



a) b)

Figure 3. Kare speaking in a) Māori and b) English. Notice how eye contact is avoided by the system while conversing in Māori, but maintained while speaking English.

(blink), etc.), which are associated with words or phrases. TalkingHead was designed for speech synchronization and thus has highly deformable lips and tongue, and is deformed parametrically. We have modified TalkingHead to produce facial expressions using the eyes and eyebrows.

5 Results

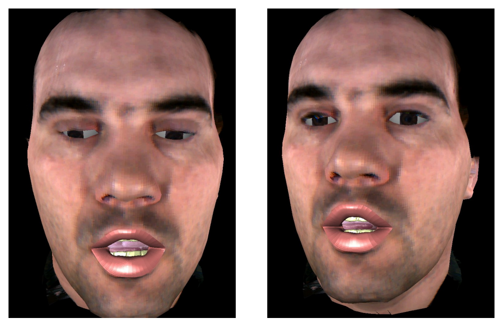
Kare is able to have a conversation (see Figure 1) with a user, albeit with a limited vocabulary. It comprehends what the user tells it, and it is able to answer questions about information the user has given it. Kare keeps track of the user with inexpensive hardware and is capable of face-to-face communication. Figure 3 contains snapshots of Kare speaking. When speaking English, eye contact is maintained with the listener. However, when speaking Māori, Kare avoids eye contact so as not to display aggression. Eye contact is avoided by looking down, looking up, or even closing the eyes; a choice made by the Id.

Figure 4 shows Kare during affirmative responses to a question. While speaking Māori the eyebrows are raised to signify a positive response. While speaking English Kare will nod its head.

Figure 5 shows negative responses to a question. For English, the head shakes side to side to convey a no. For Māori, the system stochastically chooses between a head shake and looking downward. For Māori, the gestures may also be accompanied a vocalized ‘kao’, so that the negative response is less likely to be missed.

a) b)

Figure 4. Kare giving affirmative responses in a) Māori and b) English. In Māori, the eyebrows are raised, while in English a nod is given.



a) b)

Figure 5. Kare giving a negative response in a) Māori and b) English. The system speaks ‘no’ while shaking its head for English. But in Māori the system chooses to look down while vocalizing ‘kao’.

6 Summary

Te Kaitito was designed strictly for text input and output, but because of its architecture it is quite capable of generating nonverbal behavior for an animated conversational agent. The generated nonverbal behavior is based not only on the dialogue act but also on the language used. The bilingual capabilities of both the dialogue system and the facial animation system allow for a believable conversation agent that shows potential for use in many applications such as teaching language.

Kare show great promise but it is still in its infancy. To be a truly immersive experience the system requires further work. The vocabulary of Te Kaitito is rather small and one gets tired of discussing such a small number of nouns. Also, TalkingHead currently is just a disembodied head. A character with a full body would be a better experience. The speech recognition currently only understands English. To act as a bilingual teacher, Kare should also understand Māori. As well, advanced audio processing may allow the system to teach pronunciation. The eyes of Kare are also quite simple, only seeing where the user is located. If the eyes could recognize faces, hand gestures, facial expressions and emotion, and eye gaze of the user a far superior system would result.

7 Acknowledgments

This work was partially supported by University of Otago Research Grant MFHB10, and by the NZ Foundation for Research in Science & Technology grant UOOX02. We thank Sui-Ling Ming-Wong for proofreading the text of this article.

References

[1] J. Beskow and S. McGlashan. Olga - a conversational agent with gestures.

[2] A. W. Black, P. Taylor, R. Caley, and R. Clark. The festival speech synthesis system. http://www.cstr.ed.ac.uk/projects/festival/, August 1999.

[3] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan. Embodiment in conversational interfaces: Rea. In *Proceedings of the CHI'99 Conference*, pages 520–527, Pittsburgh, PA, 1999.

[4] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Bechet, B. Douville, S. Prevost, and M. Stone. Animated conversation: Rule–based generation of facial expression gesture and spoken intonation for multiple converstaional agents.

[5] J. Cassell, H. H. Vilhjálmsson, and T. Bickmore. BEAT: the behavior expression animation toolkit. In E. Fiume, editor, *Proceedings of SIGGRAPH '01 (Los Angeles, California, August 12-17, 2001)*, Computer Graphics Proceedings, Annual Co, pages 477–486. ACM SIGGRAPH, ACM Press, August 2001.

[6] M. Cohen and D. Massaro. Modeling coarticulation in synthetic visual speech. In N. Magnenat-Thalmann and D. Thalmann, editors, *Models and Techniques in Computer Animation*, pages 139–156. Springer-Verlag, Tokyo, 1993.

[7] B. De Carolis, C. Pelachaud, I. Poggi, and F. de Rosis. Behavior planning for a reflexive agent. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, IJCAI 2001*, pages 1059–1066, Seattle, Washington, August 4–10 2001.

[8] S. de Jager, A. Knott, and I. Bayard. A DRT-based framework for presuppositions in dialogue management. In *Proceedings of the 6th workshop on the semantics and pragmatics of dialogue (EDILOG 2002)*, Edinburgh, 2002.

[9] D. DeCarlo, C. Revilla, M. Stone, and J. J. Venditti. Making discourse visible: Coding and animating conversational facial displays. In *Proceedings of Computer Animation 2002*, pages 11–16, Geneva, Switzerland, June 19-21 2002.

[10] S. Duncan and D. Fiske. *Interaction structure and strategy*. Cambridge University Press, 1985.

[11] I. Eibl-Eibesfeldt. Similarities and differences between cultures in expressive movements. In S. Weitz, editor, *Nonverbal communication*, pages 20–33. Oxford University Press, 1974.

[12] K. Grammer, W. Schiefenhövel, M. Schleidt, B. Lorenz, and I. Eibl-Eibesfeldt. Patterns on the face: the eyebrow flash in crosscultural comparison. *Ethology*, 77:279–299, 1988.

[13] S. A. King. *A Facial Model and Animation Techniques for Animated Speech*. PhD thesis, The Ohio State University, Columbus, OH, June 2001.

[14] A. Knott, I. Bayard, S. de Jager, and N. Wright. An architecture for bilingual and bidirectional nlp. In *Proceedings of the 2nd Australasian Natural Language Processing Workshop (ANLP 2002)*, 2002.

[15] J. C. Lester, J. L. Voerman, S. G. Towns, and C. B. Callaway. Cosmo: A life-like animated pedagogical agent with deictic believability.

[16] J. Metge and P. Kinloch. *Talking past each other: problems of cross-cultural communication*. Victoria University Press, Wellington, New Zealand, 1984.

[17] C. Pelachaud, N. I. Badler, and M. Steedman. Linguistic issues in facial animation. In N. Magnenat-Thalmann and D. Thalmann, editors, *Computer Animation '91*, pages 15–30. Springer-Verlag, Tokyo, 1991.

[18] C. Pelachaud and I. Poggi. Subtleties of facial expressions in embodied agents. *JVCA*, 13(5):301–312, December 2002.

[19] J. Rickel and W. L. Johnson. Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. *Applied Artificial Intelligence*, 13:343–382, 1999.

[20] A. E. Scheflen. *The Significance of Posture in Communication Systems in Communication in Face to Face Interaction*. Penguin Modern Linguistics Readings Harmondsworth: Penguin Books Ltd, 1972.

[21] The CMU Sphinx Group. CMU sphinx: Open source speech recognition. http://www.speech.cs.cmu.edu/sphinx/, 2002. Accessed Nov 15, 2002.

[22] P. Viola and M. Jones. Rapid object detection using a boosted cascade of sample features. In *Computer Vision and Pattern Recognition*, volume I, pages 511–518. IEEE Computer Society, 2001.