14

Multimodal Interaction Control

Jonas Beskow, Rolf Carlson, Jens Edlund, Björn Granström, Mattias Heldner, Anna Hjalmarsson, Gabriel Skantze

Kungl Tekniska Högskolan, Centre for Speech Technology, Stockholm, Sweden

No matter how well hidden our systems are and how well they do their magic unnoticed in the background, there are times when direct interaction between system and human is a necessity. As long as the interaction can take place unobtrusively and without techno-clutter, this is desirable. It is hard to picture a means of interaction less obtrusive and techno-cluttered than spoken communication on human terms. Spoken face-to-face communication is the most intuitive and robust form of communication between humans imaginable. In order to exploit such human spoken communication to its full potential as an interface between human and machine, we need a much better understanding of how the more human-like aspects of spoken communication work.

A crucial aspect of face-to-face conversation is what people do and what they take into consideration in order to manage the flow of the interaction. For example, participants in a conversation have to be able to identify places where it is legitimate to begin to talk, as well as to avoid interrupting their interlocutors. The ability to indicate that you want to say something, that somebody else may start talking, or that a dialog partner should refrain from doing so is of equal importance. We call this *interaction control.*

Examples of the features that play a part in interaction control include the production and perception of *auditory cues* such as intonation patterns, pauses, voice quality, and various disfluencies; *visual cues* such as gaze, nods, facial expressions, gestures, and visible articulatory movements; and *content cues* like pragmatic and semantic (in)completeness. People generally seem to use these cues in combination, and to mix them or shift between them seamlessly. By equipping spoken dialog systems with more human-like interaction control abilities, we aim to move interaction between system and human toward the intuitive and robust communication among humans.

The bulk of work on interaction control in CHIL has been focused on auditory prosodic cues, but visual cues have also been explored, and especially through the use of embodied conversational agents (ECAs) – human-like representations of a system, for example, animated talking heads that are able to interact with a user in a natural way using speech, gesture, and facial expression. ECAs are one way of leveraging

the inherent abilities that we all possess in terms of decoding information in speech, visible articulation, intonation, voice quality, facial displays, gestures, and gaze and holds the potential of improving the effectiveness, robustness, and naturalness of human-computer interaction. We have also explored the usefulness of gaze direction for interaction control purposes.

This chapter describes work along a number of lines in order to multimodally capture and mimic the flow of human-human interaction. Section 14.1 describes experiments and techniques to improve the flow of the interaction by analyzing what users say and do, and Section 14.2 concerns how multimodal output – an animated talking head and prosodically as well as temporally aware speech synthesis – can be used to achieve further improvements.

14.1 Interaction Control in Spoken Dialog Systems

As mentioned, the bulk of the work on interaction control in the CHIL project has been focused around auditory, prosodic cues. The choice was made partly because prosodic cues have several advantages from a system design point of view: They are available from the speech signal alone, and furthermore it may be possible to utilize them without using ASR. Thus, they rely less heavily on other technologies than, for example, semantic completeness or gaze tracking.

14.1.1 Silence Duration Thresholds

Current spoken dialog systems commonly detect where the user ceases speaking in order to find out where they should take their turn. The method is based on the assumption that speakers have finished what they intended to say when they become silent and that these points in time are also suitable places for the system to speak. Such *endpoint detection* triggers on a certain set amount of silence, or non-speech. The end-of-utterance detectors in current automatic speech recognition typically rely exclusively on a silence threshold somewhere between 500 and 2000 ms (sic!) (cf. [14] and references mentioned therein). The method makes sense; given that a speaker is allowed to complete what she or he intends to say, the end of the utterance is likely to coincide with silence at a place where an interlocutor might take the next turn. The method segments speech into reasonably sized units, in many cases corresponding to sentences or some sentence-like units. However, spontaneous conversational speech frequently contains silent pauses inside what we would intuitively group into turns, complete utterances, or sentence-like units, and inside what are indeed semantically coherent units, and dialog systems using silence-based endpoint detection run into problems with unfinished utterances when encountering spontaneous speech [2]. Experiments presented in [11] showed that dialog systems relying on silence-based segmentation run the risk of interrupting their users in as much as 35% of all silent pauses in the kind of speech investigated.

14.1.2 Prosodic Boundaries

It follows that something is needed in addition to silence, but what should that be? A starting point is to find out what people are able to do and distinguish in terms of suitable and unsuitable places to say something in a spoken interaction. Places where speaker changes may occur (i.e., suitable places to speak) are closely linked to the notion of prosodic boundaries. Places where the speaker is allowed to finish what she is going to say by nature coincide and co-occur with prosodic boundaries. However, it is far from obvious that every kind of prosodic boundary represents a possible place for turn-taking.

Therefore, early on in CHIL, several tests were carried out to examine the relationship between prosodic boundaries and suitable places for speaker changes. Heldner et al. [18] report two of these tests, a listening test where subjects rated the appropriateness of made-up turn-takings, and a production experiment where another group of subjects was asked to indicate suitable places for turn-takings.

The stimuli used in the listening test were made up of fragments from a seminar on speech technology given in English by a lecturer of German origin. The positions to be evaluated were selected based on manual annotations of perceived prosodic boundaries using a three-level convention (strong vs. weak vs. no prosodic boundary). Each of the stimuli in the listening test consisted of a fragment from the seminar followed by a fragment of a question from somebody in the audience. These stimuli were subsequently presented to a number of subjects in a listening experiment where the task was to rate whether the questioner enters the conversation at an appropriate place. The results of the listening test showed that the listeners generally judged the turn-takings in strong boundary positions to be more appropriate than those in weak or no boundary conditions, and that weak boundaries were slightly better than no boundaries.

A subsequent production experiment was carried out to let the subjects indicate possible places for turn-takings in the same speech material that was used in the listening experiment. Here, each trial consisted of the subjects listening to a lecturer's part of the seminar and as soon as they thought it was appropriate to take the turn, they pressed a key, the sound of the lecturer stopped, and the question, *"What about <um> could you give us some <hrm> rough idea what"* was played. Subsequently, the sequence of lecturer part and question was repeated and the subjects had to rate whether the trial was successful (discard, keep), the timing of the turn-taking on a micro level (early, OK, late), as well as the politeness of the turn-taking (rude, neutral, polite). The possible places for turn-takings indicated by our subjects were analyzed in terms of whether they occurred at a weak or a strong boundary, or not at any boundary.

The production experiment supports the findings from the listening test by showing a strong preference for turn-takings at strong boundaries, although turn-takings may also occur at certain weak and no boundaries according to our subjects. Furthermore, as nearly all strong boundaries occurring in the experiment were marked as possible turn-taking positions, it is reasonable to assume that strong boundaries

generally make up appropriate turn-taking positions, at least in this communicative situation.

Several additional studies highlighting various aspects of the relationships between prosodic boundaries and turn-taking have been performed. Some of these have been published in [18, 5, 17, 35].

14.1.3 Prosodic Cues

We opted for prosody, then, but there are still choices to be made. Previous work suggests that a number of prosodic or phonetic cues are associated with turn-yielding and thus potentially relevant for interaction control. These cues include, apart from silent pauses, phenomena such as various intonation patterns (rises, falls, down-steps, up-steps); decreases in speech rate; final lengthening; intensity patterns; centralized vowel quality; creaky voice quality; and exhalations. Note that both rises and falls have been associated with turn-yielding. These cues are typically located somewhere toward the end of the turn, although not necessarily on the final syllable [24, 25, 43, 15, 23, 28].

Similarly, there are studies suggesting that certain prosodic or phonetic cues are associated with turn-keeping, and these cues are, of course, also potentially relevant for the interaction control. They include phenomena such as glottal or vocal tract stops without audible release; a different quality of silent pauses as a result of these glottal or vocal tract closures; assimilation across the silent pause; and other "held" articulations (e.g., lengthened vowels, laterals, nasals, or fricatives) [23] [28]. In particular, level intonation patterns in the middle of the speaker's fundamental frequency range have been observed to act as turn-keeping cues in several different languages. For example, Duncan [8] reported that any pattern other than a level tone in the speaker's mid-register (a 2 2 | pattern in the Trager -Smith prosodic transcription scheme [40]) signals turn-yielding in English. Thus, the mid-level pattern acts as a turn-keeping signal, although Duncan did not use that term. Similarly, Selting [30] reported that level pitch accents before a pause are used to signal a turn-holding (or turn-keeping) in German; Koiso et al. [22] observed that flat, flat-fall, and rise-fall intonation patterns tended to co-occur with speaker holds (i.e., turn-keeping) in Japanese; and in another study on Japanese conversations Noguchi, and Den [27] reported that flat intonation at the end of pause-bounded phrases acts as an inhibitory cue for backchannels. Furthermore, in a study of final pitch accents and boundary tones in the turn-taking system of Dutch, Caspers [6] identified two intonation patterns that seem to be associated with turn-keeping: an accent lending rise followed by level high pitch (H* %) used for bridging syntactic breaks between utterances; and a filled pause with a mid-level boundary tone (M %) for bridging hesitations within syntactic constituents. However, Caspers could not find any intonation patterns clearly associated with turn-yielding. This observation led her to conjecture that turn-changing is the unmarked case and that only the wish to keep the turn needs to be marked with specific intonation patterns.

Based on this, we decided to see if this mid-level pattern could be automatically extracted and used to inform the interaction control of a spoken dialog system. A

number of requirements must be placed on such an extraction if it is to be of any practical use.

14.1.4 `/nailon/` – Online Automatic Extraction of Prosodic Cues

In order to use additional information, including prosody, to improve interaction control in practical applications such as in spoken dialog systems, the information needs to be made available to the system, which places special requirements on the analyses. First of all, in order to be useful in live situations, all processing must be performed automatically, in real time, and deliver its results with minimal latency (cf. [31]). Furthermore, the analyses must be online in the sense of relying on past and present information only, and cannot depend on any right context or look ahead. There are other technical requirements: The analyses should be sufficiently general to work for many speakers and many domains, and they should be predictable and constant in terms of memory use, processor use, and latency.

In order to meet these requirements, **`/nailon/`**, a software package for online, real-time prosodic analysis, was developed. **`/nailon/`** is based on the Snack software library [34] and provides a number of analyses. Originally, the choice of methods to include in the analyses was in part ad hoc, in part based on literature. During development and testing, most of the ad hoc methods were underpinned. A more detailed description of the workings of **`/nailon/`** can be found in [12].

In order to extract data that can be used for making decisions about interaction control – that is, pitch patterns preceding silence –**`/nailon/`** does the following, in turn:

Voice/speech activity detection (VAD/SAD). The first step is to decide whether or not there is speech. The VAD built into **`/nailon/`** is trivial and intended for testing purposes only. VAD/SAD is a research area in its own right, and for real use, methods for plugging in external VAD/SAD are provided. Note also that channel separation is not provided by **`/nailon/`**, which expects channel-separated input – one speaker per channel. The VAD/SAD is performed on the frame level, but the results may be smoothened by subsequent processes.

Extraction of pitch, intensity, and voice. Next, an ESPS extraction of pitch, intensity, and voicing [39] is done for each frame judged to be speech. This extraction is modified to run repeatedly over a small window to run at real time with a latency of one frame.

Online speaker normalization. The pitch values are normalized against an incremental model of the speaker's range to provide range relative values in terms of standard deviation [42], which in turn gives an idea of "high", "medium", and "low" pitch, from the speaker's perspective. The method has been validated in [13].

Pause detection. **`/nailon/`** flags consecutive silence of a certain length. The length used in the tests has been 300 ms, but the system works with lengths below 200 ms on a normal desktop computer. The length chosen, plus a few milliseconds for processing, constitutes the smallest amount of silence needed in the speaker channel for the system to detect a suitable place to speak. It should be compared to the 500-2000 ms used by most current systems.

Convex hull extraction. Whenever silence of a certain length is found, a modified convex hull extraction [26] is used to find the last syllable-like entity – pseudo-psyllable or psyllable – preceding the silence. The use of convex hulls was inspired by Nick Campbell and colleagues, who used convex hulls for automatic classification of phrase final tones in the (JST/CREST ESP Project) [20].

Finally, the decision of whether or not a given silence is a suitable place to speak is made not by **`/nailon/`**, but by the interaction control software, which may be trained or rule-driven. In either case, the normalized pitch pattern over the last psyllable preceding the silence is the input used for such a decision.

The data extracted by **`/nailon/`** have been used to perform interaction control studies. In [11] it was shown that the number of incorrect turn-taking decisions can be reduced substantially by combining standard silence-based endpoint detection with an automatic classification of intonation patterns. In the process, it is also possible to decrease the length of the required silence without any loss in performance. This can be used to make a conversational computer more responsive by allowing it to reply faster without simultaneously making it more obtrusive.

Automatically classified level intonation patterns in the middle of the speaker's fundamental frequency range were found to act as turn-keeping cues, and may thus be used to avoid interrupting human interlocutors with high precision. Although there are several observations of the function of these mid-level intonation patterns, to our knowledge, they have never been used to avoid interrupting users of spoken dialog systems before.

14.1.5 Interaction Model

Before we turn to the production, or output, side of spoken human-machine interaction, we need a model of its flow that caters to simultaneous speech, multiple speakers, etc. For this purpose, we have developed an interaction model [10] (see Fig. 14.1) that can be viewed as an extended version of the AVTA system [21] but differs in that it models the situation subjectively and separately for each participant, while it lacks a combined, objective "God's view" of the interaction. Although we have not exploited this property within the CHIL project, it is worth noting that this subjectiveness is useful for modeling interaction under transmission latency, since it can capture the fact that the effect of round-trip latency is perceived differently by different subjects.

The model is computationally simple yet powerful. It consists of three parts: a state derived directly from each participant's speech activity, a state derived from the speech activity of all participants, and events representing changes in these states.

The first state (SPEECH/SILENCE) continuously models speech/nonspeech as a binary state on a per-participant level. At any given point in time, each participant may be either speaking or not speaking. The only input the model takes is speech/nonspeech decisions from each participant's VAD.

The second part of the model is a four-way decision of the communicative state (SELF/OTHER/NONE/BOTH), again repeated for each participant. These states are derived from the SPEECH/SILENCE state of each participant. From participant P's point

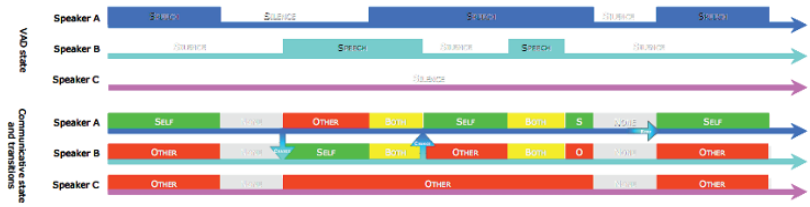

Fig. 14.1. Model of spoken interaction flow.

of view, the state is NONE if none of the participants is speaking. It is SELF if P is speaking but no one else is. If one or more other participants are speaking and P is silent, it is OTHER; finally, if both P and some other participant are speaking, it is BOTH.

Finally, the model includes transitions from one communicative state to another for each participant. If P is in state NONE and someone else starts speaking, P goes from NONE to OTHER and the participant who started speaking goes from NONE to SELF.

The system is modeled as any other participant in the model. At any given point in time, the current state (SELF/OTHER/NONE/BOTH) from the system's point of view can be read. If the system is silent and has something to say, either it may wait for the NONE state to appear, or it may decide to barge in. In making these decisions, other factors than the state of the interaction model can be weighted. Typically, factors such as how urgently the system wants to say something (e.g., if the message is that the fire alarm has been set off, the system should disregard any politeness considerations); what type of utterance the system wants to make (feedback and backchannels, for example, may be easier to fit into other people's speech and also cause less frustration if slightly mismatched); and, of course, cues such as intonation and gaze, as discussed above. Other contextual features matter as well, for example, the emotional state of the current speaker(s) (see Chapter 10).

14.1.6 Other Modalities: Eye Gaze

As widely noted in the literature and above, speaker changes in conversation correlate with a great many things that are not prosody. Among the more prominent ones is eye gaze. In the CHIL project, a study using a Tobii x50 eye tracker (see Fig. 14.2 and 14.3) corroborates the statement that eye gaze behavior is relevant for interaction control. The study shows that people tend to look more at each other when they listen than when they speak, as has been reported in many previous studies. The results also show that interlocutors tend to glance at each other around the time of speaker changes, presumably to ensure understanding as well as a smooth speaker change [19].

Fig. 14.2. Experimental setup.

Fig. 14.3. Tobii x50 eye tracker.

14.2 Multimodal Output and Interaction Control

Many of the interaction experiments in CHIL incorporate the use of an animated talking head. We now turn to provide a general overview of the talking head animation system used in this research, and specifically the generation of expressive visual speech animation, which is a prerequisite for reaching the type of conversation on human terms that we have targeted in the project. An experiment to assess the speech intelligibility gain provided by a talking head when used in combination with the targeted audio system that is described in Chapter 13 is also included, as it relates to interaction control in that the aim is not to disrupt the flow of a meeting, for example.

14.2.1 Animated Talking Head

Our talking head is based on the MPEG-4 Facial Animation standard [29]. It is a textured 3D model of a male face comprised of approximately 15,000 polygons. The mesh has been parameterized to allow for realistic deformation, using a framework based around a combination of professional 3D modeling tools and in-house custom algorithms, and a flexible animation engine.

In order to render visual speech movements, we employ a data-driven methodology. We start from a time-aligned transcription of the speech to be synthesized. The time-aligned transcription can be obtained from a text-to-speech system, if we are synchronizing with synthetic speech, or it can be produced by a phoneme recognizer (as in the Synface system [3]) or a phonetic aligner [33]. Next, an articulatory control model is applied to convert the time-aligned phonetic transcription of the utterance into control parameter trajectories to drive the articulation of the talking head model. In order to produce convincing and smooth articulation, the articulatory control model has to account for coarticulation, which refers to the way in which the realization of a phonetic segment is influenced by neighboring segments.

14.2.2 Expressive Speech

The articulatory control model has been trained to reproduce the facial movements of a real speaker. While previous work on visual speech synthesis typically has been aimed at modeling neutral pronunciation, this model is also capable of synthesizing expressive speech. We used an opto-electronic motion tracking system to collect a multimodal corpus of acted emotional speech. The system allows the dynamics of emotional facial expressions to be captured by registering the 3D coordinates of a number of reflective markers at a rate of 60 frames/second. For this study, our speaker, a male native Swedish amateur actor, was instructed to produce 75 short sentences with the six emotions of happiness, sadness, surprise, disgust, fear, and anger, plus neutral, yielding 7 times 75 recorded utterances. A total of 29 IR-sensitive markers were attached to the speaker's face, four of which were used as reference markers (on the ears and on the forehead). The marker setup largely corresponds to MPEG-4 feature point (FP) configuration.

The recorded corpus with expressive speech was then used to train articulatory control models based on PCA analysis [7] for the five emotions happy, angry, surprised, sad, and neutral. To provide suitable training data, the recorded marker positions were first encoded as MPEG-4 FAPs, and the audio track was phonetically labeled using a forced alignment system. The resulting models can later be used to synthesize articulatory movements for novel arbitrary Swedish speech, thereby modeling expression and articulation in an integrated fashion. Figure 14.4 shows snapshots from resulting animations of the same utterance for the four emotions happy, angry, surprised, and sad. See [4] for further details.

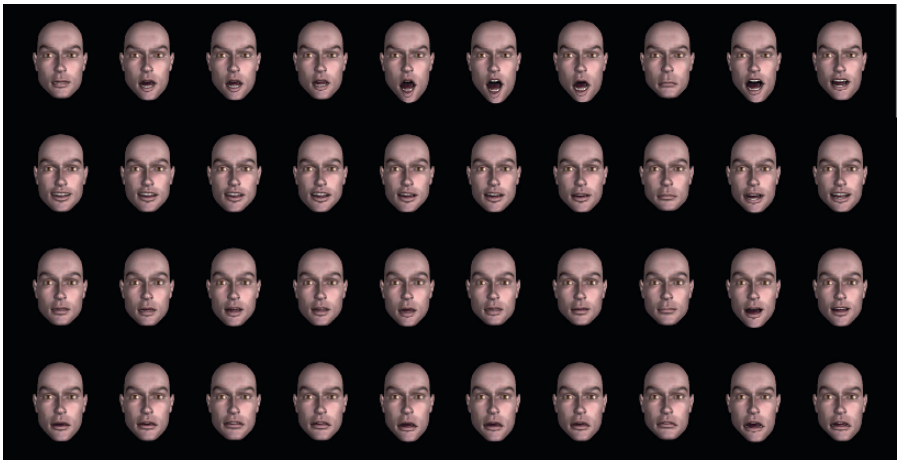

Fig. 14.4. Snapshots taken every 0.1 s from the animation of the fragment "jag ska köpa..." ("I will buy..."), synthesized with four different expressive speech models: happy (top row), angry (second row), surprised (third row), and sad (bottom row).

14.2.3 Unobtrusive Speech

While properly synchronized and articulated visual speech synthesis is an important property of animated talking agents for multimodal interaction that improves realism, it also adds to the intelligibility of the speech output [32]. This property may be exploited for spoken interaction in noisy environments. It may also be used, for example, in quiet meeting-room scenarios to deliver spoken messages to individuals without disturbing the meeting, since it allows the volume to be kept to a minimum. A talking head combined with directed audio makes it possible to achieve even less disturbance for meeting partners.

An experiment was conducted to test how the intelligibility of speech is affected by the listener's position relative to the audio beam and to measure the possible augmentation of intelligibility that the use of a talking head can provide for this type of application. Two different listening positions were tested: one where the subject was positioned in the audio beam (0°) and one where the loudspeaker was rotated 45° away from the subject. For each position, both an audio-only condition and a condition where a talking head was displayed on the screen were evaluated.

The distance between the subject's head and the loudspeaker was approximately 80 cm. To better simulate a real environment, no specific means were taken to ensure that the subject remained still during the experiment. Due to problems with sound reflection, and in order to minimize disturbance outside the audio beam, a low-output sound level was used. The task consisted of listening to nonsense VCV (vowel-consonant-vowel) words and identifying the consonant in each word. The seven consonants to be identified were [f, s, m, n, k, p, t] uttered in an [a_a] context, produced by a male speech synthesis voice. Each of these seven VCV words occurred

a total of four times in each condition, and the presentation order was randomized. The order in which the different conditions were presented was also rotated to avoid having possible learning effects affect the overall result. The answering sheet consisted of a forced choice among the seven consonants in the test.

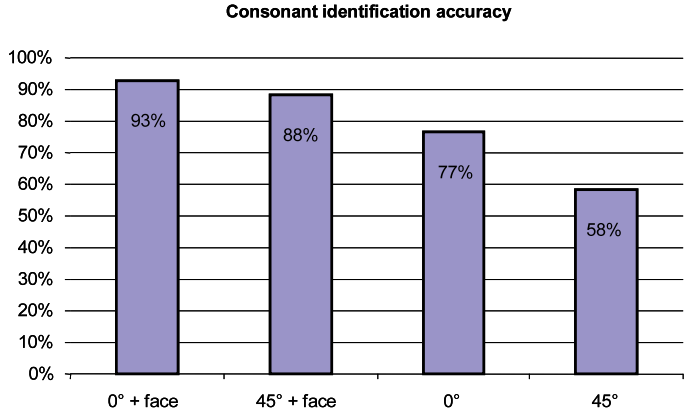

Fig. 14.5. Recognition rates for nonsense words with targeted audio, with and without talking head support.

The results of the test are shown in Fig. 14.5. The highest recognition accuracy, 93%, was obtained for the condition where the subjects were positioned in the audio beam and the talking head was used (0° + face condition). For the 45° + face condition, the recognition rate was 88%. When only the sound was available, 77% recognition accuracy was reached for the 0° condition, and 58% for the 45° condition. The result of this last condition (audio-only, 45°) was significantly ($p < 0.05$) from the other three conditions. The differences in accuracy for the 0° condition, between having access to a talking face or not (i.e., 0°+audio-only vs. 0°+face), was also significant ($p < 0.05$). For further details, see [38].

14.2.4 MushyPeek – An Experimental Framework

Another aspect affecting the decision, from the system's point of view, of whether to speak or remain silent, is that it is not a binary choice. The system can also opt to make "feelers" to see whether it is given the floor or not. It may do this vocally, for example, by "clearing its throat" softly, but it may also use gestures in an animated talking head representing the system. In [10], we present MushyPeek, an experimental framework inspired by [16], in which listeners can hear the speakers' real voices while watching what they are told to be graphic representations of the speakers and their gestures on monitors. The framework can be seen as a special case of the *transformed social interaction* discussed by Bailenson et al. [1]. In MushyPeek,

the participants are placed in separate rooms, and each participant is equipped with a headset connected to a Voice-over-IP call (http://www.skype.com/). On both sides, the call is enhanced with SynFace [3] – a lip-synchronized animated talking head representing each participant. As both talking heads represent real persons (the participants), we refer to them as *avatars* in the following. This basic setup constitutes the communicative backbone of the framework. In addition, the framework contains experiment-specific components for voice activity detection (VAD), interaction modeling and control, gesture realization, and logging. All components communicate over TCP/IP connections. The framework is symmetrical in that both participants have the same setup. The general layout is shown in Fig. 14.6.

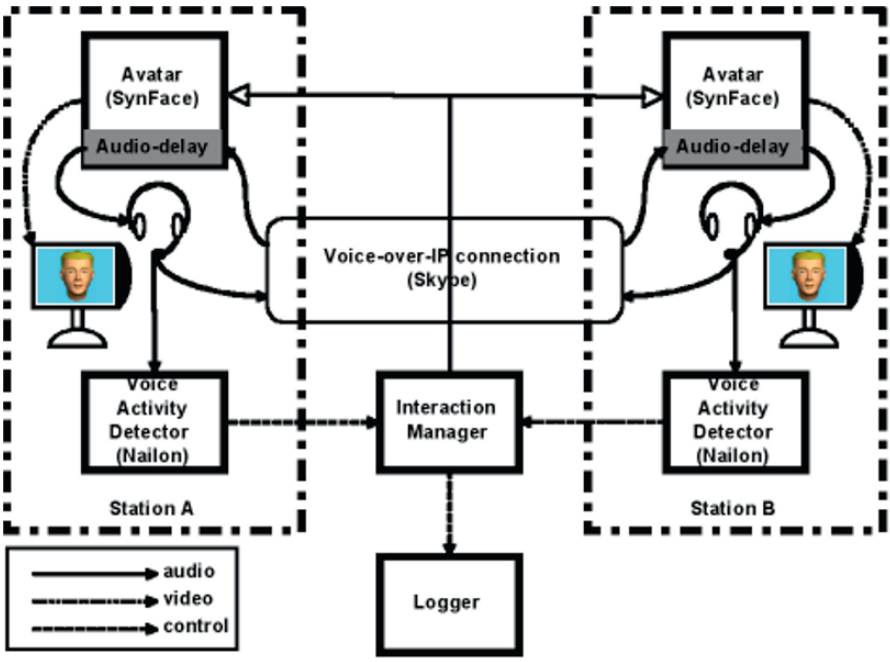

Fig. 14.6. The MushyPeek experimental framework.

A first experiment in MushyPeek using the interaction model and /nailon/ prosodic analysis, together with turn-taking gestures in an animated talking head, showed that it is indeed possible to unobtrusively affect people's willingness to release the floor in this manner [9].

14.2.5 Short Feedback Utterances

A final aspect that is crucial to creating an unobtrusive and smooth flow in human-machine interaction has to do with the manner in which the system says things. Within CHIL, KTH has investigated the use of small responsive backchannels [41] as well as unobtrusive and responsive feedback and clarification requests [37, 36]. These aspects have also been implemented in many demo applications, such as *the Hummer*, which provides backchannels such as "uh-huh" and "OK" at appropriate places, and an animated talking head doing narration while taking the listener's level of attention into consideration. The Hummer has been demonstrated in plenum at several conferences, and the narrating talking head has been demonstrated at technology days and at IST Helsinki 2007.

References

1. J. N. Bailenson, A. C. Beall, J. Loomis, J. Blascovich, and M. Turk. Transformed social interaction: Decoupling representation from behavior and form in collaborative virtual environments. *Presence: Teleoperators & Virtual Environments*, 13(4):428–441, Aug. 2004.
2. L. Bell, J. Boye, and J. Gustafson. Real-time handling of fragmented utterances. In *Proceedings of NAACL Workshop on Adaptation in Dialogue Systems*, pages 2–8. Carnegie Mellon University, Pittsburgh, PA, 2001.
3. J. Beskow, I. Karlsson, J. Kewley, and G. Salvi. *SYNFACE – A Talking Head Telephone for the Hearing-Impaired*, pages 1178–1186. Springer-Verlag, New York, 2004.
4. J. Beskow and M. Nordenberg. Data-driven synthesis of expressive visual speech using an MPEG-4 talking head. In *Proceedings of Interspeech 2005*, Lisbon, Sept. 2005.
5. R. Carlson, J. Hirschberg, and M. Swerts. Cues to upcoming Swedish prosodic boundaries: Subjective judgment studies and acoustic correlates. *Speech Communication*, 46:326–333, 2005.
6. J. Caspers. Local speech melody as a limiting factor in the turn-taking system in Dutch. *Journal of Phonetics*, 31:251–276, 2003.
7. M. M. Cohen and D. W. Massaro. *Modelling Coarticulation in Synthetic Visual Speech*, pages 139–156. Springer Verlag, Tokyo, 1993.
8. S. Duncan, Jr. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2):283–292, 1972.
9. J. Edlund and J. Beskow. Pushy versus meek – using avatars to influence turn-taking behaviour. In *Proceedings of Interspeech 2007 (ICSLP)*, pages 682–685, Antwerp, Belgium, 2007.
10. J. Edlund, J. Beskow, and M. Heldner. Mushypeek – an experiment framework for controlled investigation of human-human interaction control behaviour. In *TMH-QPSR Vol. 50: Proceedings of Fonetik 2007*, pages 65–68, Stockholm, 2007.
11. J. Edlund and M. Heldner. Exploring prosody in interaction control. *Phonetica*, 62(2-4):215–226, 2005.
12. J. Edlund and M. Heldner. /nailon/ – software for online analysis of prosody. In *Proceedings of the Ninth International Conference on Spoken Language Processing (Interspeech 2006 - ICSLP)*, Pittsburgh, PA, 2006.

13. J. Edlund and M. Heldner. Underpinning /nailon/: Automatic estimation of pitch range and speaker relative pitch. In C. Müller, editor, *Speaker Classification*. Springer/LNAI, forthcoming.
14. L. Ferrer, E. Shriberg, and A. Stolcke. Is the speaker done yet? faster and more accurate end-of-utterance detection using prosody in human-computer dialog. In *Proceedings of the Seventh International Conference on Spoken Language Processing (ICSLP 2002)*, volume 3, pages 2061–2064, Denver, CO, 2002.
15. C. E. Ford and S. A. Thompson. Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns. In E. Ochs, E. A. Schegloff, and S. A. Thompson, editors, *Interaction and grammar*, pages 134–184. Cambridge University Press, Cambridge, 1996.
16. J. Gratch, A. Okhmatovskaia, F. Lamothe, S. Marsella, M. Morales, R. J. van der Werf, and L.-P. Morency. Virtual rapport. In *Proceedings of 6th International Conference on Intelligent Virtual Agents*, Marina del Rey, CA, US, 2006.
17. M. Heldner, J. Edlund, and T. Björkenstam. Automatically extracted f0 features as acoustic correlates of prosodic boundaries. In *Proceedings of Fonetik 2004*, pages 52–55. Department of Linguistics, Stockholm University, 2004.
18. M. Heldner, J. Edlund, and R. Carlson. Interruption impossible. In M. Horne and G. Bruce, editors, *Nordic Prosody: Proceedings of the IXth Conference, Lund 2004*, pages 97–105. Peter Lang, Frankfurt am Main, 2006.
19. V. Hugot. *Eye gaze analysis in human-human communication*. Master thesis, KTH Speech, Music and Hearing, 2007.
20. C. T. Ishi. Perceptually-related f0 parameters for automatic classification of phrase final tones. *IEICE Transactions on Information and Systems*, E88D(3):481–488, 2005.
21. J. Jaffe and S. Feldstein. *Rythms of dialogue*. Personality and Psychopathology. Academic Press, New York, 1970.
22. H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den. An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs. *Language and Speech*, 41(3-4):295–321, 1998.
23. J. K. Local and J. Kelly. Projection and "silences": Notes on phonetic and conversational structure. *Human Studies*, 9:185–204, 1986.
24. J. K. Local, J. Kelly, and W. H. G. Wells. Towards a phonology of conversation: turn-taking in Tyneside English. *Journal of Linguistics*, 22(2):411–437, 1986.
25. J. K. Local, W. H. G. Wells, and M. Sebba. Phonology for conversation: Phonetic aspects of turn delimitation in London Jamaican. *Journal of Pragmatics*, 9(2-3):309–330, 1985.
26. P. Mermelstein. Automatic segmentation of speech into syllabic units. *Journal of the Acoustical Society of America*, 58(4):880–883, 1975.
27. H. Noguchi and Y. Den. Prosody-based detection of the context of backchannel responses. In *Proceedings of the Fifth International Conference on Spoken Language Processing (ICSLP'98)*, pages 487–490, Sydney, Australia, 1998.
28. R. Ogden. Turn transition, creak and glottal stop in finnish talk-in-interaction. *Journal of the International Phonetic Association*, 31(1):139–152, 2001.
29. I. S. Pandzic and R. Forchheimer. *MPEG-4 Facial Animation – the Standard, Implementation and Applications*. John Wiley & Sons, Chichester, 2002.
30. M. Selting. On the interplay of syntax and prosody in the constitution of turn-constructional units and turns in conversation. *Pragmatics*, 6:357–388, 1996.
31. E. Shriberg and A. Stolcke. Direct modeling of prosody: An overview of applications in automatic speech processing. In *Proceedings of Speech Prosody 2004*, pages 575–582. Nara, Japan, 2004.

32. C. Siciliano, G. Williams, J. Beskow, and A. Faulkner. Evaluation of a multilingual synthetic talking face as a communication aid for the hearing impaired. In *Proc of ICPhS, XV International Conference of Phonetic Sciences*, pages 131–134, Barcelona, Aug. 2003.
33. K. Sjölander and M. Heldner. Word level precision of the nalign automatic segmentation algorithm. In *Proceedings of the XVIIth Swedish Phonetics Conference, Fonetik 2004*, pages 116–119, Stockholm University, may 2004.
34. K. Sjölander. The Snack sound toolkit, 1997. `http://www.speech.kth.se/snack/`.
35. G. Skantze and J. Edlund. Robust interpretation in the higgins spoken dialogue system. In *COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction*, Norwich, UK, 2004.
36. G. Skantze, D. House, and J. Edlund. Grounding and prosody in dialog. In *Working Papers 52: Proceedings of Fonetik 2006*, pages 117–120. Lund University, Centre for Languages & Literature, Department of Linguistics & Phonetics, Lund, Sweden, 2006.
37. G. Skantze, D. House, and J. Edlund. User responses to prosodic variation on fragmentary grounding utterances in dialogue. In *Proceedings Interspeech 2006*, pages 2002–2005, Pittsburgh, PA, 2006.
38. G. Svanfeldt and D. Olszewski. Perception experiment combining a parametric loudspeaker and a synthetic talking head. In *Proceedings of Interspeech*, pages 1721–1724, 2005.
39. D. Talkin. A robust algorithm for pitch tracking (rapt). In B. Kleijn and K. Paliwal, editors, *Speech Coding and Synthesis*, pages 495–518. Elsevier, New York, NY, 1995.
40. G. L. Trager and H. L. Smith. *An Outline of English Structure*. American Council of Learned Societies, Washinton, DC, 1957.
41. Å. Wallers, J. Edlund, and G. Skantze. The effects of prosodic features on the interpretation of synthesised backchannels. In E. André, L. Dybkjaer, W. Minker, H. Neumann, and M. Weber, editors, *Proceedings of Perception and Interactive Technologies*, pages 183–187. Springer, Kloster Irsee, Germany, 2006.
42. B. Welford. Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4(3):419–420, 1962.
43. B. Wells and S. MacFarlane. Prosody as an interactional resource: Turn projection and overlap. *Language and Speech*, 41(3-4):265–294, 1998.