

Multimodal Action
Professor Benjamin Weiss
Quality and Usability Lab
Technische Universität Berlin
Turn Taking

(Refer Slide Time: 00:17)

Multimodal Action

- Outline:**
- Introduction: Verbal and nonverbal information
 - Gestures
 - Posture
 - Gaze
 - Voice
 - Space
 - Turn taking
 - Emotion
 - Social relationship
 - Summary



Turn-taking is a special topic.

(Refer Slide Time: 00:21)

Turn taking

- Turn-taking as interaction control**
- Organization of the allocation of the right to participate in interaction (Sacks et al., 1974)
 - There is usually no pause in face-to-face interaction
 - There is often a small overlap (cultural dependent)
 - People organize turn-taking
 - At a transition relevant point, the speaker (turn holder) signals either
 - Selects a new speaker
 - Holds the turn
 - Gives the turn (also to oneself)



This is about coordination of people who are engaged in a conversation. So when people speak to each other, turn-taking is the topic about finding out and determining who has the floor, who is the person who actually speaks.

Of course in a conversation of two people or more in a group everybody is active; even the listeners. They constantly or repeatedly signal back information. This is called back-channeling or feedback. And this is about giving signals that they are still listening; they are still following something like go on, or here I am, I am still, you still have my attention.

This can be verbal like uh huh, Ok or some small comments but they usually are non-verbal, this kind of back-channeling like uh um, or even like some facial expressions or head movements and so on.

The important thing is that this back-channeling does not disturb and interrupt the Speaker who has the floor. It signals that they are attentive and it can even affect the talker a little bit.

That means, if I have some emotional response to what the speaker says or if I am a little bit in doubt, if this is really true I can signal this to the speaker and he or she will for example, follow this and elaborate a little bit more or react on my back-channeling. But overall the speaker still has the floor and is speaking.

So turn-taking is about the change of who is speaking, who has the floor. And this is signaled and coordinated at so-called transition relevant points. So for example, if I am speaking here and I am finishing my utterance or my point that I want to make, everybody who is engaged in the conversation can foresee this, can predict this.

And in the traditional, transition relevant point when I, I am going down in my voice a little bit low pitch, I am going to be a little bit slower and at this point there are actually 3 options. One option is that I determine the next person to have the floor. For example by actually, explicitly stating this like Fredrick, what do you think about that?

But of course there are other options. Maybe I do not want to give away the floor, maybe I want to continue although I have done my point and finished it. I will then signal that I still want to continue and want to hold the floor. I do this, for example by not going slower or down in pitch. I may even inhale visibly and audibly so that the listeners, my dialog partners know that I want to continue.

The third option is that I am giving the floor away. I am giving the floor for free for everybody to grab it. Typically the active listeners would have already announced by some non-verbal signals that they are interested in taking the floor but if nobody does, does it, I can even take the floor again myself.

So how does this look like?

(Refer Slide Time: 03:48)

Turn taking

Turn-taking as interaction control

- Signals include gaze, intonation, gesture, or speech (give the turn)
- Intonation: final syllable lengthening, lower pitch vs. yield turn (higher pitch, filled pauses)
- Gesture: less gesturing with hands
- Gaze: look at interlocutor (vs. avoiding eye contact)
- Mutual break: gaze is broken by the listener and the turn is taken (more often, more frequent in smooth interaction)
- Mutual hold: gaze is held by the listener and the turn is taken
- Verbal: "What do you think, Paul?" (vs. inhaling)
- Signals to request turns include: inhaling, starting to speak, starting hand gestures, and body movements (e.g. straight posture)
- Feedback is not part of turn taking
- Nod, head tilt, facial expression (smile, surprise), acoustics (mhm, "really?")
- Functions: acknowledgement, attention/involvement signaling, problem/clarification signaling



These signals that indicate I giving away the floor, I have already done some examples in voice. They are non-verbal and multimodal and that means I can also decrease a little bit in my gesturing. I start looking at the other people, gazing at them. These are all signals that now there is a transition relevant point and I am giving away the floor.

So what do listeners do? If they want to attract the attention, want to indicate that they want to have the floor they also show this by signaling. For example by leaning forward or making themselves tall or by audibly inhaling or starting to gesture a little bit.

So with gaze, we have for example two patterns that are well-known, the mutual break and the mutual hold. The mutual break is the more typical one or the smoother one. That means if I have the floor and at a transition relevant point I end my turn then I lock gaze with listener who wants to have the floor and then he or she breaks the gaze again and starts talking. This is a rather smooth pattern.

And other one, the mutual hold, this is little bit different. In this case the listener does not break the gaze pattern but still has locked eyes and starts talking. From a little study there was shown that this is indicating a more non-smooth overlap or non-smooth turn-taking.

The interesting thing about turn-taking is not only that is multimodal but also that is really smooth. And this means typically we predict when it is going to be to happen.

And therefore it means that the actual breaks or pauses that we have during turn-taking are rather small, sometimes they are just a small part of overlap which is not disturbing anybody from talking or there is even a rather small break or pause or silence.

If there is too much silence then everybody would consider this is a little bit awkward and non-smooth. So turn-taking is this action control which is predicted and prepared by everybody who is actively engaged in the conversation and it is signaled in a multimodal way.

(Refer Slide Time: 06:30)

Turn taking

- Turn-taking as interaction control
- Current goal: Implementation of better (natural) interaction management
 - Incremental ASR (anticipate user statements)
 - Faster reaction
 - 'interrupt' user, natural turn-taking
 - Requires also incremental TTS and DM (Schlangen, 2005)



And as you can imagine in current human computer interfaces this is not the case.

Typical speech related systems; they do not predict when the users will have finished his or her utterance. Typically, we have a kind of push the talk system which means that I as user push a button, real button or soft button on a Smartphone and I indicate by this when I am talking to the machine.

Or there is machine or the computer is waiting for stretch of silence of some hundreds of milliseconds and then starts processing my input trying to interpret this then producing the system response.

And this is rather small and more like a walkie-talkie style, not like a natural way to interact with each other by voice as we are used to that. And therefore current research is going in the direction of enabling systems to also act like a human by providing proper turn-taking signals.

For this we need a new architecture. We need a so-called incremental architecture because the system has constantly to listen to the user, trying to interpret what the intention of my speech is and prepare proper reactions. In order to people get ready when the turn transition relevant point comes and the turn-taking takes place.

(Refer Slide Time: 08:01)

Multimodal Action

- Outline:**
- Introduction: Verbal and nonverbal information
 - Gestures
 - Posture
 - Gaze
 - Voice
 - Space
 - Turn taking
 - Emotion
 - Social relationship
 - Summary



Quite related to turn-taking in a matter that it is multimodal is emotion otherwise it is not that related.

(Refer Slide Time: 08:11)

Emotion

- Expression of emotion is audio-visual
- A lot of nonverbal information is revealed in interaction by gestures, mimics, gaze, positioning, posture, and voice
 - Social status, age, gender, personality, likeability, agreement, interest
- Like language, nonverbal interaction is mostly dependent on culture (not like extreme basic emotions like „fear“)



Emotions are signaled of course by posture but mostly, the most research that we have done is concerning facial expressions and voice.

With emotions we can assume that they are so-called basic or pure options. This is of course debated but we can consider that there are some emotions or categories of emotions which are so strong and widespread that they are culturally independent and universal like fear for life or pure hatred or strong joy.

So one of the pioneers for this was actually Charles Darwin. We know from

(Refer Slide Time: 09:03)

Emotion

- Emotions are a long term subject of research (cf. e.g. Darwin, 1872: The expression of emotions in man and animals)
- There are rare cases, when humans express naturally pure emotions (fear of death, match win etc.), but in every day situations, emotions are not extreme and not pure, i.e. not socially controlled
- Emotions are apparent in facial expressions, e.g. those 6 are considered universal



facial expressions that it is quite easy to determine, interpret which kind of emotion has which facial expression.

You see here in the so-called 6 basic emotions at least provided by some researchers. The actual number is not that certain but from left to right, the 6 emotions are joy, sadness, anger, surprise, disgust and the last one is fear.

This seems to be quite obvious. But we have similar expressions in voice.

(Refer Slide Time: 09:46)

Emotion

- These six "basic" emotions are also recognizable in speech:
 - Joy: e.g. high pitch, high 2.3. spectral formant (smiling), raising pitch contours
 - Anger: e.g. high pitch, many strong stressed syllables, rough voice, falling pitch contours, precise pronunciation
 - Fear: e.g. slightly raised pitch, high speech tempo, less often raised pitch an utterance ends
 - Disgust: e.g. slow tempo, long stressed syllables, less pitch variation
 - Sadness: e.g. less pitch variation, less pitch accents, slow tempo, long pauses
 - Boredom (to replace surprise): e.g. rhythmic stress pattern, slow tempo, waved pitch contour
- Scherer et al., 2013
<http://www.speech.synchronic.de/>
- Basic emotions are considered universal
 - Showing emotions (apart from extreme one) is cultural dependent
 - Example of a multimodal synthesis



Here is a small list provided by Scherer which kind of so-called prosodic or voice-related parameters and features there are to express or to perceive basic emotions in voice. I also have some audio examples of German actors trying to act such an emotion. I will only playback few of them.

So joy for example is quite distinctive because joy usually comes together with a spreading of

(Refer Slide Time: 10:23)

Emotion

- These six "basic" emotions are also recognizable in speech:
 - Joy: e.g. high pitch, high 2.,3. spectral formant (smiling), raising pitch contours
 - Anger: e.g. high pitch, many strong stressed syllables, rough voice, falling pitch contours, precise pronunciation
 - Fear: e.g. slightly raised pitch, high speech tempo, less often raised pitch an utterance ends
 - Disgust: e.g. slow tempo, long stressed syllables, less pitch variation
 - Sadness: e.g. less pitch variation, less pitch accents, slow tempo, long pauses
 - Boredom (to replace surprise): e.g. rhythmic stress pattern, slow tempo, waved pitch contour
- Scherer et al., 2003
<http://emosamples.syntheticspeech.de/>
- Basic emotions are considered universal
 - Showing emotions (apart from extreme one) is cultural dependent
 - Example of a multimodal synthesis



the lips. And this actually affects the frequency energy distribution. So you may remember from an earlier video, the so-called formants, the residence, the resonance frequencies in the mouth and these are affected by the spreading of the lips.

But also joy is typically exhibited by a, or signaled by a high pitch and rather strong melody. So raising up and down movements

(Refer Slide Time: 10:58)

Emotion


- These six "basic" emotions are also recognizable in speech:
 - Joy: e.g. high pitch, high 2.,3. spectral formant (smiling), raising pitch contours
 - Anger: e.g. high pitch, many strong stressed syllables, rough voice, falling pitch contours, precise pronunciation
 - Fear: e.g. slightly raised pitch, high speech tempo, less often raised pitch an utterance ends
 - Disgust: e.g. slow tempo, long stressed syllables, less pitch variation
 - Sadness: e.g. less pitch variation, less pitch accents, slow tempo, long pauses
 - Boredom (to replace surprise): e.g. rhythmic stress pattern, slow tempo, waved pitch contour
- Scherer et al., 2003
<http://emosamples.syntheticspeech.de/>
- Basic emotions are considered universal
 - Showing emotions (apart from extreme one) is cultural dependent
 - Example of a multimodal synthesis

of pitch.

Another example is anger. Anger, for example has really strong stressed syllables and also a falling intonation contour. Here is the example of German actor.

(Refer Slide Time: 11:17)

Emotion

- These six "basic" emotions are also recognizable in speech:
 - Joy: e.g. high pitch, high 2.,3. spectral formant (smiling), raising pitch contours
 - Anger: e.g. high pitch, many strong stressed syllables, rough voice, falling pitch contours, precise pronunciation
 - Fear: e.g. slightly raised pitch, high speech tempo, less often raised pitch an utterance ends
 - Disgust: e.g. slow tempo, long stressed syllables, less pitch variation
 - Sadness: e.g. less pitch variation, less pitch accents, slow tempo, long pauses
 - Boredom (to replace surprise): e.g. rhythmic stress pattern, slow tempo, waved pitch contour
- Scherer et al., 2003
- Basic emotions are considered universal
 - Showing emotions (apart from extreme one) is cultural dependent
 - Example of a multimodal synthesis
- <http://emosamples.syntheticspeech.de/>
- 


11:15 demo start

11:21 demo end

And as you have noticed it is also quite loud. Fear on the other hand typically has a soft intensity, soft voice; that means it is not that loud. However, it has a high pitch and a usually higher tempo. Disgust on the other hand has a slow tempo and long-stressed syllables and less pitch variation, and here is an example.

(Refer Slide Time: 11:47)

Emotion

- These six "basic" emotions are also recognizable in speech:
 - Joy: e.g. high pitch, high 2.,3. spectral formant (smiling), raising pitch contours
 - Anger: e.g. high pitch, many strong stressed syllables, rough voice, falling pitch contours, precise pronunciation
 - Fear: e.g. slightly raised pitch, high speech tempo, less often raised pitch an utterance ends
 - Disgust: e.g. slow tempo, long stressed syllables, less pitch variation
 - Sadness: e.g. less pitch variation, less pitch accents, slow tempo, long pauses
 - Boredom (to replace surprise): e.g. rhythmic stress pattern, slow tempo, waved pitch contour
- Scherer et al., 2003
- Basic emotions are considered universal
 - Showing emotions (apart from extreme one) is cultural dependent
 - Example of a multimodal synthesis
- <http://emosamples.syntheticspeech.de/>
- 

11:47 demo start



11:53 demo end

Maybe you have noticed there is another issue. She sounds like having a really closure in the throat and just like, as she would have liked to prevent to eat something that is really not tasty. But this is just one prototypical typical way to signal disgust. There are other ways.

Sadness has less pitch variations. It is very monotonic and also very slight pitch accent. So not as stressed as with anger for example. It often comes together with a slow tempo and long pauses. Here is

(Refer Slide Time: 12:32)

Emotion

- These six "basic" emotions are also recognizable in speech:
 - Joy: e.g. high pitch, high 2.,3. spectral formant (smiling), raising pitch contours
 - Anger: e.g. high pitch, many strong stressed syllables, rough voice, falling pitch contours, precise pronunciation
 - Fear: e.g. slightly raised pitch, high speech tempo, less often raised pitch an utterance ends
 - Disgust: e.g. slow tempo, long stressed syllables, less pitch variation
 - Sadness: e.g. less pitch variation, less pitch accents, slow tempo, long pauses
 - Boredom (to replace surprise): e.g. rhythmic stress pattern, slow tempo, 
- Scherer et al., 2003
- Basic emotions are considered universal
 - Showing emotions (apart from extreme one) is cultural dependent
 - Example of a multimodal synthesis 
- <http://emosamples.syntheticspeech.de/>

an example.

12:33 demo start

12:37 demo end

Maybe you have heard that this particular actor, this particular way of signaling sadness also results in, in a fading of voicing in the end. That means as if she is physically not able to have the strength to properly pronounce and articulate.

The last example is boredom. And this is actually just one of the basic emotions that we defined. So this is typically not in the set of the six basic emotions. It is rather similar to sadness in a way that it is slow and low in pitch. But here you find a certain difference in, for example physical strength that is signaled.

13:28 demo start

13:33 demo end


In contrast to sadness she starts really high in pitch and she has a certain way of rhythmic pattern and this is quite typical, this rhythmic pattern of showing that it is apparently something that comes again and again is therefore really boring.

I have on the slides a link to the, to the right where you have a nice example, a nice collection of modern and older text to speech synthesis voices which are able to produce emotions or effective states. Check it out if you are interested in that.

The most important thing about emotions, not only that it is multimodally signaled by, for example facial expressions and voice, but also that of despite being universal they are cultural differences. At least when we are able to speak we have some control about our body. This means the cultural preferences of showing or not showing emotions takes place here.

Now I will show you a small example how in a virtual person emotions are synthesized. This is again I think about 10 years old but it was nicely the different modalities that can be used for producing and synthesizing emotions.

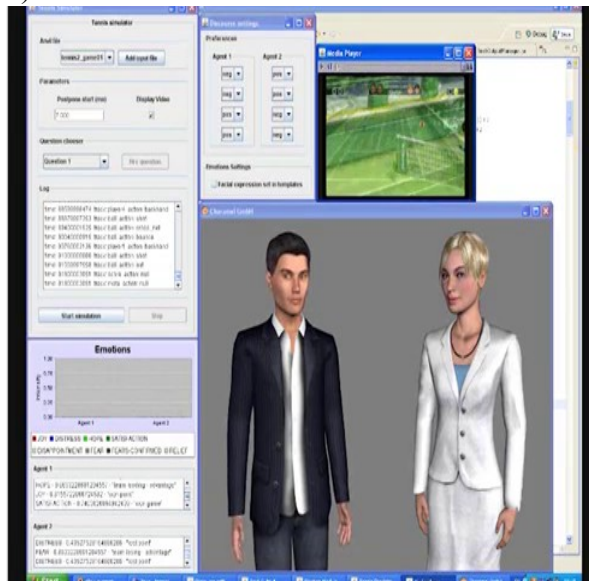
(Refer Slide Time: 14:58)



*Agents are engaged
in dialogues*

15:00 demo start

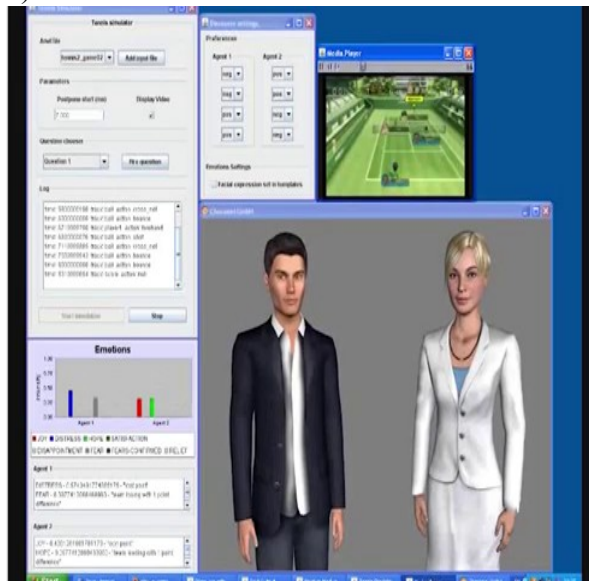
(Refer Slide Time: 15:01)



(Refer Slide Time: 15:15)



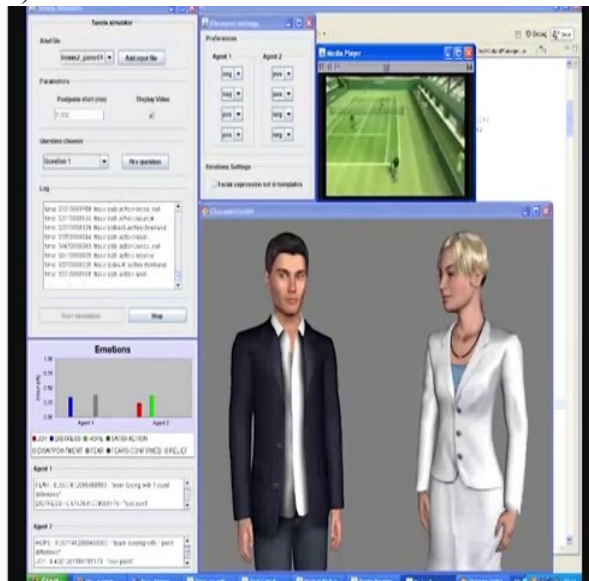
(Refer Slide Time: 15:18)



(Refer Slide Time: 15:41)



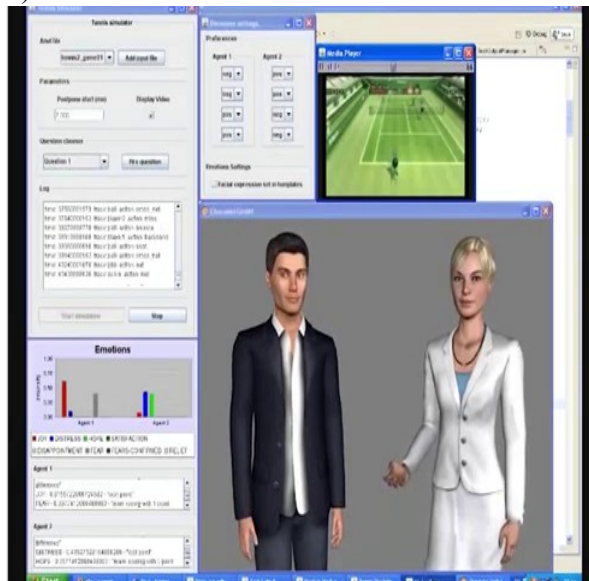
(Refer Slide Time: 15:45)



(Refer Slide Time: 16:00)



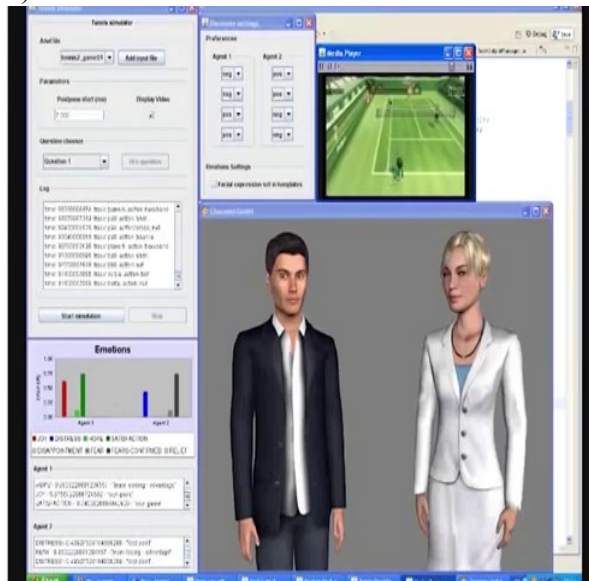
(Refer Slide Time: 16:03)



(Refer Slide Time: 16:17)



(Refer Slide Time: 16:20)



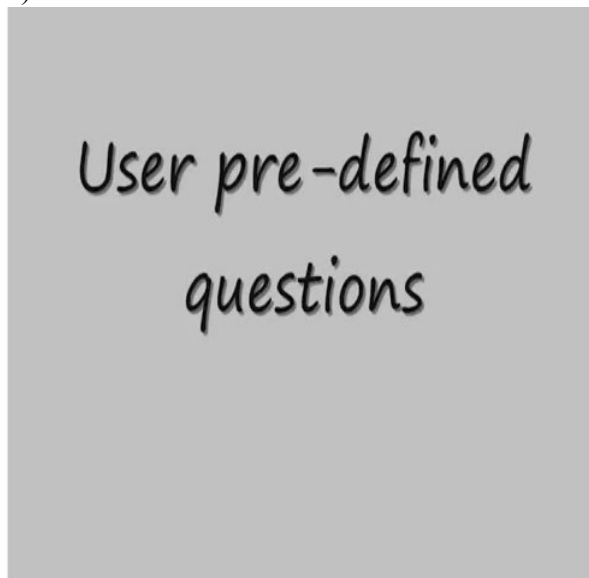
(Refer Slide Time: 16:28)



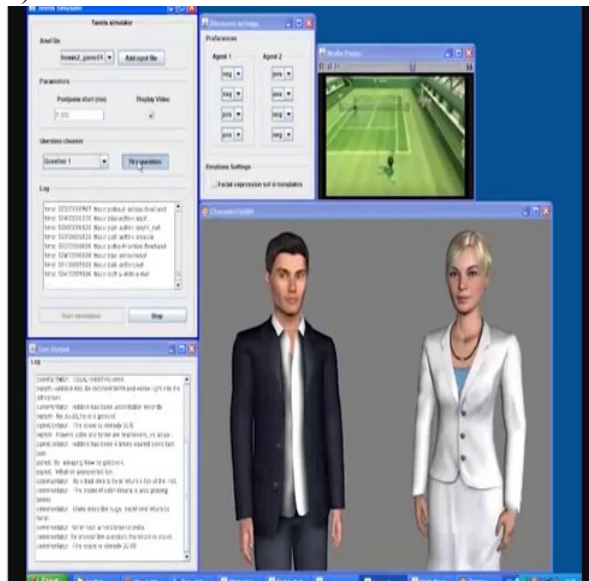
(Refer Slide Time: 16:30)



(Refer Slide Time: 16:37)



(Refer Slide Time: 16:40)



16:54 demo end