Modality–specific Resources: Extension

Pilar Manchón, David Ávila, Antonio Ávila

Distribution: Public

TALK
Talk and Look: Tools for Ambient Linguistic Knowledge
IST-507802 Deliverable 3.3

February 15, 2007

Project funded by the European Community
under the Sixth Framework Programme for
Research and Technological Development

The deliverable identification sheet is to be found on the reverse of this page.

Project ref. no.	IST-507802
Project acronym	TALK
Project full title	Talk and Look: Tools for Ambient Linguistic Knowledge
Instrument	STREP
Thematic Priority	Information Society Technologies
Start date / duration	01 January 2004 / 36 Months

Security	Public
Contractual date of delivery	M36 = December 2006
Actual date of delivery	February 15, 2007
Deliverable number	3.3
Deliverable title	Modality–specific Resources: Extension
Type	Report
Status & version	Draft Version date: February 15, 2007
Number of pages	12 (excluding front matter)
Contributing WP	??
WP/Task responsible	??
Other contributors	??
Author(s)	Pilar Manchón, David Ávila, Antonio Ávila
EC Project Officer	Evangelia Markidou
Keywords	

The partners in TALK are:

Saarland University	USAAR
University of Edinburgh HCRC	UEDIN
University of Gothenburg	UGOT
University of Cambridge	UCAM
University of Seville	USE
Deutches Forschungszentrum fur Künstliche Intelligenz	DFKI
Linguamatics	LING
BMW Forschung und Technik GmbH	BMW
Robert Bosch GmbH	BOSCH

For copies of reports, updates on project activities and other TALK-related information, contact:

The TALK Project Co-ordinator
Prof. Manfred Pinkal
Computerlinguistik
Fachrichtung 4.7 Allgemeine Linguistik
Postfach 15 11 50
66041 Saarbrücken, Germany
pinkal@coli.uni-sb.de
Phone +49 (681) 302-4343 - Fax +49 (681) 302-4351

Copies of reports and other material can also be accessed via the project's administration homepage, `http://www.talk-project.org`

Contents

Contents

Chapter 1

Introduction

1.1 Introduction: The MIMUS Talking Head

In order to provide a more complete multimodal showcase, USE has implemented a Talking Head that complements the previous implementations, and endows the system with additional features and communicative intensity.

The overall purpose of the MIMUS system is to become a practical and valuable tool in the smart home scenario, and more in particular, an everyday tool for the specific focus group proposed.

It is therefore understandable that with that objective in mind, different sub–objectives gain importance: human–like interaction must be not only efficient, but may and/or should also include additional human features. In order to endow the system with sufficient capabilities to fulfill these requirements, the MIMUS system has been furnished with a talking head that complements the system's personality, and confers an appearance of human–like communication on the interaction.

The literature [8] [1] illustrates how different experiments show that:

- Computers are indeed social actors
- The users' conduct is quite different when interacting with a virtual character as opposed to when they interact with a faceless computer.
- The overall user satisfaction is greater when interacting with a virtual character

MIMUS seeks to be a clear example of user–centered design, and with the user always in mind, the MIMUS talking head has been integrated into the main system architecture.

Chapter 2

The MIMUS Talking Head

2.1 What's in a name: Sebastian, Ambrosio or Albert?

In the same way as we all have specific preferences and personal affinities, a virtual character with which a user will be often interacting, and which, to some degree, she will depend upon, must either have a very neutral and non–transcending personality, or a concurrent one.

Although a rather good looking and smiling face would have gained more positive reactions initially, a rather dry and not–so–pleasant character was chosen in order to generate a more noticeable impact. With this character, USE intended to take advantage of some cultural stereotypes to invest the system with an aura of credibility: a toffee–nosed butler always of service that suggests efficiency and properness.

In order to create a coherent character, the name is key. Different names have been selected according to the main interaction language. "Ambrosio" is a typical Spanish name for a butler, as well as "Sebastian" in English or "Albert" in German. Although the veracity of this "cultural feeling" has not been formally confirmed as far as we know, casual surveys support the choice.

Endowing the character with a name has a manifold purpose:

- Personalization
- Personification
- Voice activation

Ambrosio will remain inactive until called for duty (voice activation); each user may name their personal assistant as they wish (Personalization); and they will address the system at personal level, reinforcing the sense of human–like communication (Personification).

2.2 Talking head design and implementation

The latest technical advancements in speech synthesis, graphical interfaces and hardware have made possible a new generation of virtual–character–based interfaces:

- Interactive multimodal interfaces with very high speech quality and lip synchronization
- More human–like behaviour at perceptual level
- More information redundancy and natural cues to support interaction

The virtual head has been implemented in 3D to allow for more natural and realistic gestures and movements. The graphical engine used is OGRE [6], a powerful, free and easy to use tool.

2.3 3D Facial Animation

There are four main aspects to be taken into account when designing a 3D talking face:

1. Modeling
2. Expressiveness
3. Animation
4. Texture

Modeling: The modeling methodology chosen is based on the facial muscular structure [5], which determines the basic modeling lines and areas that will in turn allow for the generation of facial expressions.

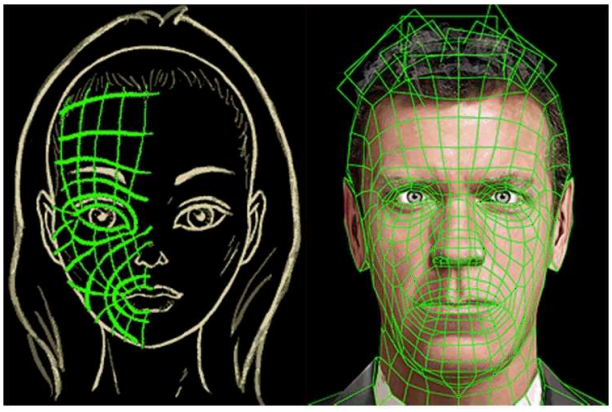

Figure 2.1: Modeling

Expressiveness: In a 3D real–time application, facial expressions are generated by means of pre–defined poses. Once each expression is modeled, the 3D vertex variation for each pose is recorded separately,

so that several expressions can be simultaneously generated. This is a widely used method called lineal interpolation animation. Each expression can be distilled into a set of numbers (vectors), where each number represents the degree of variation from the neutral pose. The generation of simultaneous facial expressions is equal to the weighed vectorial sum of the expressions to be combined.

Figure 2.2: Expressiveness

Animation: The animation is achieved throughout a skeleton system: each bone has an impact on the neighboring vertexes. Each vertex has an associated list where each bone has an associated value ranging from 0 to 1. The value '0' indicates that this bone has no influence on that particular vertex. '1' however indicates the bone has full control over this vertex. All intermediate values indicate the relative level of influence of the bone on the vertex. The bone motion can then trigger the proportional motion of the associated vertexes in the facial mesh, achieving a more organic effect.

Texture: One of the main trade–offs in 3D real–time graphical applications is quality versus performance: the greater the level of detail, the lower the system performance. The quality of the light is a key factor to achieve high quality, minimizing the impact on performance. The complexity of light and its associated problems such as surface scattering drain the graphic card. So, given the rather static nature of a talking head, integrating the light in the texture is a smart move. Several pictures in different angles can then be composed into a 3D model with the light already integrated into the texture.This method is essential for picture–like realism.

2.4 Architecture

As in [7], the system consists of four different subsystems:

1. Input
2. Synchronization

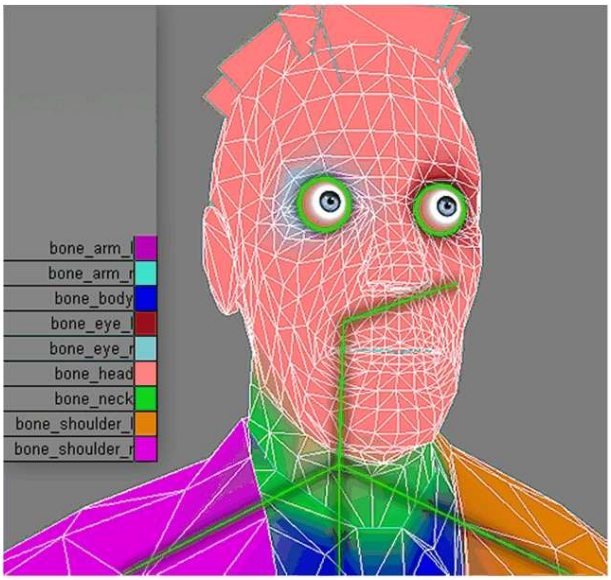

Figure 2.3: Animation

3. Speech synthesis
4. Face management

In Figure 2.5, a general overview of the subsystems is provided.
The key element here is “Synchronization”, which coordinates the synthesizer output (speech) with the lip movement handled by the face manager. In addition to this, it is important to provide an easy way to control the heads behavior while keeping it natural.

2.5 TTS Integration

As previously stated, one of the critical issues in the implementation of a talking head is the integration and synchronization of the head with the speech output. In order to do this, it is essential that the synthesizer can provide real time information about the synthesis and, more precisely, about the phonemes generated.
The current talking head is integrated with Loquendo, a high quality commercial synthesizer that launches the information about the phonemes as asynchronous events, which allows for lip synchronization

1. **Visemes** A viseme can be defined as the graphical counterpart of a phoneme, i.e., the lip movement and shape of a talking head when pronouncing a certain phoneme. Since there is no one–to–one correspondence, several phonemes can be associated with the same viseme. Several intermediate steps are necessary:
   (a) First, the TTS coding standard must be identified. In this case, SAMPA [10]
   (b) Then, the set of plausible visemes must be determined, and the phoneme–viseme mapping established. The current implementation follows the SAPI table [4]. According to SAPI, only 22 visemes are necessary to generate any dialogue.

Figure 2.4: Texture

(c) Finally, each viseme is associated with a particular virtual model pose.

This is all accomplished by means of XML configuration files, which in turn provides several advantages:

(a) Steps 1 and 2 make the system TTS–independent.

(b) Step 3 makes the model application–independent.

This way, each phoneme generated by the TTS can be associated with a virtual model pose, so that they can be activated as the string of phonemes is received, which confers a sense of naturalness on the virtual character. Following previous implementations [3] and to avoid abrupt and unnatural viseme transitions, interpolation curves have been used, providing naturalness and fluency.

2. **Interpolation curves** The amplitude 'a' of a viseme is the percentage or impact it has on the lips, ranging from '0' to '1', where '0' would imply the viseme is imperceptible, and '1' would be a full range viseme. When a viseme is activated, the representation of the function amplitude–time is as shown in Figure 2.7

   Unlike in [3], the interpolation function is based on the sin function, which provides a smooth curve. The parameters such as rise time and down time in the preceding curve are proportional to the viseme duration. In other words, longer visemes such as vowels would have a softer transition than a shorter one. The results prove that this strategy renders much better and more natural results. The transition between visemes is therefore progressive, and more natural.

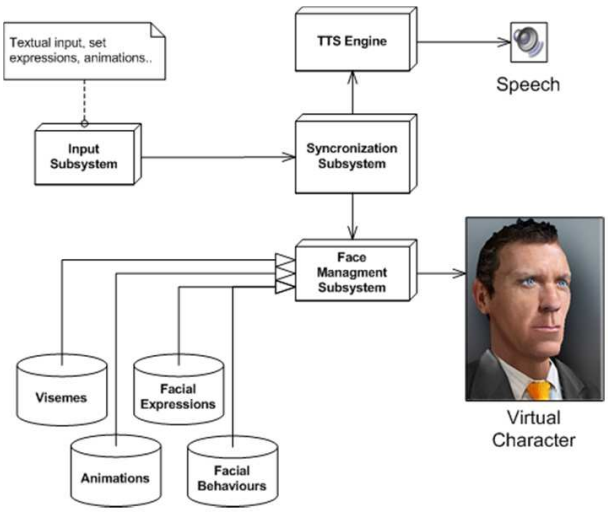

Figure 2.5: System Architecture

2.6 Expressiveness

In order to be coherent with the overall purpose of the system, a virtual character must be invested with the highest degree of credibility. According to the literature [9], gestures and expressions are almost or quite as important as speech itself. In any case, it reinforces the overall communicative act.
The first step is therefore to determine the different behaviour layers, and establish which gestures are conscious and which unconscious.

1. **Behaviour layers** Different behaviour layers can be modeled, each of which acts independently. The combination of two or more produces a full spectrum of possibilities:

   - Moods: Are mutually exclusive, that is, the virtual character can only be in one mood at a time.
     - Basic expressions according to Ekman [2].
   - Dialogue: It is associated with all gestures and/or motions that have to do with the dialogue and the communication channel. It does not require external information. It includes but is not limited to attention–related issues, turn taking, emphasis, etc.
     - Speaker attention
     - Pronunciation–related gestures
     - ...
   - Motions: All face or body movements that do not depend directly on the discourse.
     - Conscious motions
       * Shaking
       * Nodding
       * ...

VISEME	Fonema
SP_VISEME_0	Silence
SP_VISEME_1	ae, ax, ah
SP_VISEME_2	aa
SP_VISEME_3	ao
SP_VISEME_4	ey, eh, uh
SP_VISEME_5	er
SP_VISEME_6	y, iy, ih, ix
SP_VISEME_7	w, uw
SP_VISEME_8	ow
SP_VISEME_9	aw
SP_VISEME_10	oy
SP_VISEME_11	ay
SP_VISEME_12	h
SP_VISEME_13	r
SP_VISEME_14	l
SP_VISEME_15	s, z
SP_VISEME_16	sh, ch, jh, zh
SP_VISEME_17	th, dh
SP_VISEME_18	f, v
SP_VISEME_19	d, t, n
SP_VISEME_20	k, g, ng
SP_VISEME_21	p, b, m

Figure 2.6: Visemes and Phonemes correspondence

  - Unconscious motions
    * Breathing
    * Blinking
    * ...

2. **Basic expressions** According to the literature [2], there are six high–level basic expressions:

- Happiness.
- Sadness.
- Disgust.
- Anger.
- Fear.

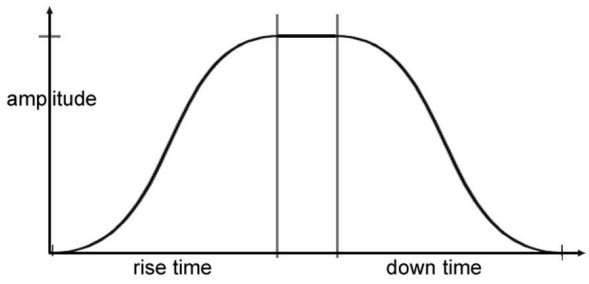

Figure 2.7: Interpolation curve of a viseme

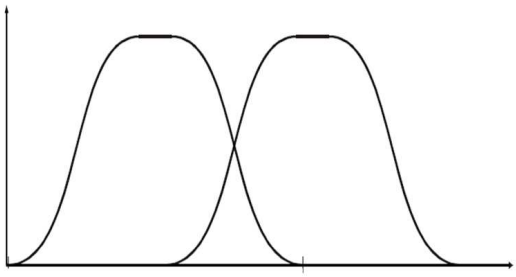

Figure 2.8: Two visemes sequence

- Surprise.

Given that in the current scenario "Disgust" was not deemed to be useful, this expression has been substituted by "Doubt", to reinforce the human–computer communication, especially given the limitations of state–of–the–art speech recognition nowadays.

These expressions suffice to model the general facial behaviour. The mood is therefore defined by the amplitude of the expression. Two expressions can be combined with specific weights, allowing in this way for the smooth transition between expressions by means of sin interpolations curves, as with visemes transitions.

2.7 The Talking head integration within the system

As it is the case with the rest of the MIMUS constituents, the talking head is an OAA agent which offers a number of solvables:

- sayText(Text): It tells the talking head what to say, and does not have any output.
  - Text. Text to say
- setLanguage(Language). It changes the TTS language, and does not have any output.
  - Language. The specific language. Only three languages are available at the moment:

    * English
    * Spanish. Default Language
    * German

- setVoicePitch(Pitch). It modifies the fundamental frequency (pitch), and does not have any output.
  - Pitch. Voice pitch.
- setVoiceSpeed(Speed).It modifies the voice rate (wm – words–per–minute), and does not have any output.
  - Speed. Speed, rate, expressed in words per minute.
- setExpression(ExpressionName): It sets the talking head expression, and does not have any output.
  - ExpressionName. Expression identifier. The following moods are available:
    * Anger
    * Doubt
    * Fear
    * Happiness
    * Idle
    * Sadness
    * Surprise
- playMotion(MotionName): It instructs the talking head to play a motion, such as nodding or shaking. It does not have any output.
  - MotionName. Motion identifier. Only these motions are currently available:
    * nod
    * shake
- wakeUp: It wakes the talking head up. If it is not asleep, the solvable does not have any effect. It has neither output nor parameters.
- sleep: It sets the talking head to sleep. If it is already asleep, the solvable does not have any effect. It has neither output nor parameters.

The dialogue manager controls the talking head, and sends the appropriate commands depending of the dialogue needs. As discussed previously, the talking head is synchronized with the synthesizer. In any case, throughout the dialogue the dialogue manager may see it fit to reinforce the communication channel with gestures and expressions, which may or not imply synthesized utterances. The head may just nod to acknowledge a command, without uttering words.

2.8 Conclusions and Future Work

In this extension to the previous deliverable (D3.3), the implementation and integration of a 3D talking head in MIMUS has been described.

We concluded from the experiments that a human–like talking head would have a significant positive impact on the subjects' perception and willingness to use the system. Since MIMUS is primarily a practical system, an effort has been made to extend the scope of the project to include the talking head.

Although no formal evaluation of the system has taken place, MIMUS has already been presented successfully in different forums, and as expected, "Ambrosio" has always made quite an impression, making the system more appealing to use and approachable.

Bibliography

[1] Bill Byrne. conversational' isn't always what you think it is. *Speech Technology*, 8(4):16, August 2003.

[2] P. Ekman and W.V. Friesen. Manual for the facial action coding system. Technical report, Consulting Psychologist Press, Inc., Palo Alto, CA, 1977.

[3] Ariel Fischer Jörn Ostermann, Mark Beutnagel and Yao Wang. Integration of talking heads and text–to–speech synthesizers fot visual tts. In *International Conference of Speech and Language Processing*, Sydney, Australia, 1998. ICSLP98.

[4] Microsoft. Sapi microsoft speech api. http://www.microsoft.com/speech/default.mspx., 1994.

[5] Arnold Moreaux. *Anatomie Artistique de l'Homme*. Maloine, Paris, 2nd edition, 1990.

[6] OGRE. Open source graphics engine. www.ogre3d.org, 2006.

[7] M. Gattass P. S. Lucena and L. Velho. Expressive talking heads: A study on speech and facial expression in virtual characters. *Scientia*, 13(2):1–12, 2002.

[8] B. Reeves and C. Nass. *The Media Equation*. CSLI–Cambridge University Press, 1998.

[9] S. Abe K. Mase T. Yonezawa, N. Suzuki and K. Kogure. Cross–modal coordination of expressive strength between voice and gesture for personified media. In *International Conference on Multimodal Interfaces (ICMI06)*, pages 43–50, Banff, Canada, November 2006. ICMI'06, ACM Press.

[10] J. C. Wells. *Handbook of Standards and Resources for Spoken Language Systems*, chapter SAMPA computer readable phonetic alphabet, pages Part IV, section B. Mouton de Gruyter, Berlin and New York, 1997.