Lexical semantics and auditory presentation in virtual storytelling

Lexical Semantics and Auditory Presentation in Virtual Storytelling

Research paper for the ICAD05 workshop "Combining Speech and Sound in the User Interface"

Minhua Ma and Paul Mc Kevitt

School of Computing & Intelligent Systems,
Faculty of Engineering,
University of Ulster, Magee Campus
Derry/Londonderry, BT48 7JL
Northern Ireland.
{m.ma,p.mckevitt@ulster.ac.uk}

ABSTRACT

Audio presentation is an important modality in virtual storytelling. In this paper we present our work on audio presentation in our intelligent multimodal storytelling system, CONFUCIUS, which automatically generates 3D animation speech, and non-speech audio from natural language sentences. We provide an overview of the system and describe speech and non-speech audio in virtual storytelling by using linguistic approaches. We discuss several issues in auditory display, such as its relation to verb and adjective ontology, concepts and modalities, and media allocation. Finally we conclude that introducing linguistic knowledge provides more intelligent virtual storytelling, especially in audio presentation.

1. INTRODUCTION

Multimodal virtual reality applications such as online games, virtual environments, and virtual storytelling, are more and more demanding the ability to render not only visual but audio scenes. A goal of our work is to create rich auditory environments that can augment 3D animation in virtual storytelling. This paper presents our work on auditory presentation in our intelligent multimodal storytelling system, CONFUCIUS, and proposes a linguistically-based approach to transform written language into multimodal presentations, including speech and non-speech sounds. We believe that integrating linguistic knowledge can achieve more intelligent multimodal storytelling which best employs different modalities to present stories, and the methodology we proposed here can serve as a framework for researchers in auditory display.

First, in section 2 we introduce background of this study--the intelligent multimodal storytelling system, CONFUCIUS, and review various nonspeech audio that could be used in virtual storytelling. Next in section 3, a linguistically-based approach for auditory presentation is proposed. We discuss several issues of this approach such as the verb/adjective ontology for audio semantics. Then we describe the auditory presentation of CONFUCIUS in section 4. Finally, section 5 compares our work to related research on virtual storytelling and summarizes the work with a discussion of possible future work.

2. BACKGROUND AND PREVIOUS WORK

We are developing an intelligent multimedia storytelling interpretation and presentation system called CONFUCIUS. It automatically generates 3D animation and speech from natural language sentences as shown in Figure 1. The input of CONFUCIUS is sentences taken from children's stories like "Alice in Wonderland" or play scripts for children. CONFUCIUS' multimodal output include 3D animation with speech and nonspeech audio, and a presentation agent—Merlin the narrator. Our work on virtual storytelling so far focussed on generating virtual human animation and speech with particular emphasis on how to use visual and audio presentation to cover more verb classes.

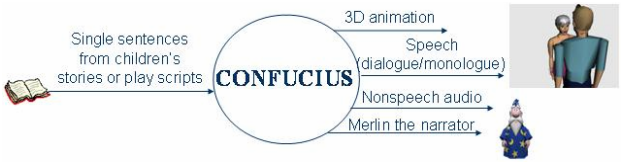

Figure 1. Input and output of CONFUCIUS

Figure 2 shows the architecture of CONFUCIUS. The dashed part in the figure is the knowledge base including language knowledge (lexicons and a syntax parser) which is used in the Natural Language Processing (NLP) module, and visual/audio knowledge such as 3D models of characters, props, and animations of actions, which encapsulate their nonspeech auditory information and are used in the animation engine. The surface transformer takes natural language sentences as input and manipulates surface text. The NLP module uses language knowledge to parse sentences and analyse their semantics. The media allocator then generates an XML-based specification of the desired multimodal presentation and assigns content to three different media: animation and nonspeech audio, characters' speech, and narration, e.g. it sends the parts bracketed in quotation marks near a communication verb to the text-to-speech engine. The animation engine takes semantic representation and use visual knowledge to generate 3D animations. The animation engine and Text-to-Speech (TTS) operate in parallel. Their outputs are combined in the synchronizing module, which outputs a holistic 3D virtual world including animation and speech in VRML format. Finally the narration integration module integrates the VRML file with

the presentation agent, Merlin the Narrator, to complete a multimedia story presentation.

We use VRML to model 3D objects and virtual characters in our story world. VRML spares efforts on media coordination since its *Sound node* is responsible for describing how sound is positioned and spatially presented within a scene. It can also describe a sound that will fade away at a specified distance from the `Sound` node by `ProximitySensor`. This facility is useful in presenting non-speech sound effects in storytelling. It enables us to encapsulate sound effects within object models, e.g. to encapsulate the engine hum within a car model and hence locate the sound at a certain point where the car is. The Sound node brings the power to imbue a scene with ambient background noise or music, as well.

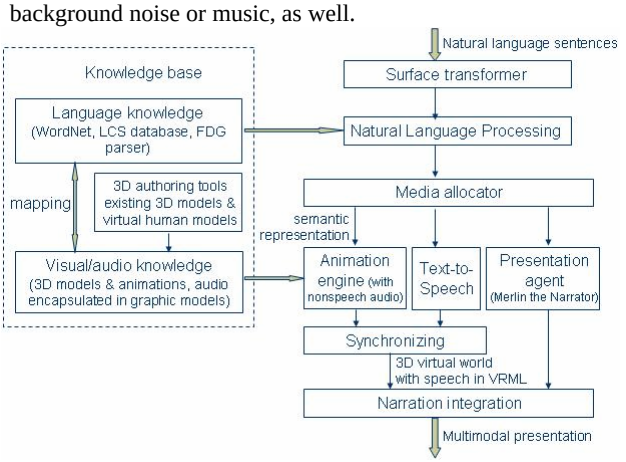

Figure. 2. Architecture of CONFUCIUS

2.1. Nonspeech audio

The use of nonspeech audio to convey information in intelligent multimedia presentation is referred to in the human factors literature as auditory display. Besides basic advantages such as reducing visual clutter, avoiding visual overload, and not requiring focused attention, auditory displays have other benefits: detection times for auditory stimuli were shorter than for visual stimuli [1], and short-term memory for some auditory information is superior to the short-term memory for visual information.

Current research in the use of nonspeech audio can generally be divided into two approaches. The first focuses on developing the theory and applications of specific techniques of auditory display. The techniques of *auditory icons*, *earcons*, *sonification*, and *music synthesis* have dominated this line of research and are discussed in detail here below. The second line of research examines the design of audio-only interfaces--much of this work is concerned with making GUIs accessible to visually-impaired users, or explores the ways in which sound might be used to extend the existing visual interface.

Auditory icons are caricatures of naturally occurring sounds which convey information by analogy with everyday events [1]. Gaver motivates this technique by questioning our basic notion of listening. In Gaver's view when we listen to sounds in our daily lives we do not hear the pitch or the duration of the sound. Rather, we hear the source of the sound and the attributes of the source. He refers to two types of listening: *musical listening* and *everyday listening*. Everyday listening includes common sounds such as the sound of pouring water, tearing paper, a car engine, or telephone ring. People tend to identify these sounds in terms of the object and events that caused them, describing their sensory qualities only when they could not identify the source events, i.e. we do not seem to hear sounds, but instead the sources of sound. Supposing that everyday listening is often the dominant mode of hearing sounds, Gaver argues that auditory displays should be built using real-world sounds. Theoretically, the advantage of auditory icons seems to be in the intuitiveness of the mapping between sounds and their meaning. Auditory icons accompanying daily life events are a major source of nonspeech audio in CONFUCIUS. Certainly the intuitiveness of this approach to auditory display will result in more vivid story presentation.

Earcons are melodic sounds, typically consisting of a small number of notes, with musical pitch relations (Gaver 1989,1). They relate to computer objects, events, operations, or interactions by virtue of a learned mapping from experience. The basic idea of earcons is that by taking advantage of sound dimensions, such as pitch, timbre, and rhythm, information can be communicated to the user efficiently. Of the four basic techniques for auditory display, earcons have been used in the largest number of computer applications. The simplest earcons are auditory alarms and warning sounds such as incoming e-mail notification, program error, and low battery alarm on mobile phones. The effectiveness of an earcon-based auditory display depends on how well the sounds are designed.

Sonification is the technique of translating multi-dimensional data directly into sound dimensions. Typically, sound parameters such as amplitude, frequency, attack time, timbre, and spatial location are used to represent system variables (Bly et al. 1987). The goal is synthesizing and translating data from one modality, perhaps a spatial or visual one, to the auditory modality. Sonification has been widely applied to a wealth of different domains: synthesized sound used as an aid to data visualisation (especially abstract quantitative data), for program comprehension, and monitoring performance of parallel programs.

In *synthesized music*, sounds are interpreted for consonance, rhythm, melodic content, and hence are able to present more advanced information such as emotional content. Computer-based music composition initiated in the mid 1950s when Lejaren Hillier and Leonard Isaacson conducted their first experiments with computer generated music on the ILLIAC computer at the University of Illinois. They employed both a rule-based system utilising strict counterpoint and a probabilistic method based on Markoff chains. The recent history of automated music and computers is densely populated with examples based on various theoretical rules from music theory and mathematics. Developments in such theories have added to the repertoire of intellectual technologies applicable to the computer. Amongst these are the Serial music techniques, the application of music grammars, sonification of fractals, and chaos equations, and connectionist pattern recognition techniques based on work in neuro-psychology and artificial intelligence.

Figure 3 illustrates the four types of nonspeech audio described above and their common features. Auditory icons and earcons are small pieces of audio clips (audio icons); sonification and synthesized music can generate audio from other modal data; and earcons and synthesized music are melodic sound.

3. A LINGUISTIC APPROACH FOR CONVERTING NATURAL LANGUAGE TO AUDITORY DISPLAY

In human commonsense knowledge, there is a natural mapping between audio and objects, events, status, and emotions. We

discuss the relations between lexical semantics and the audio modality in this section.

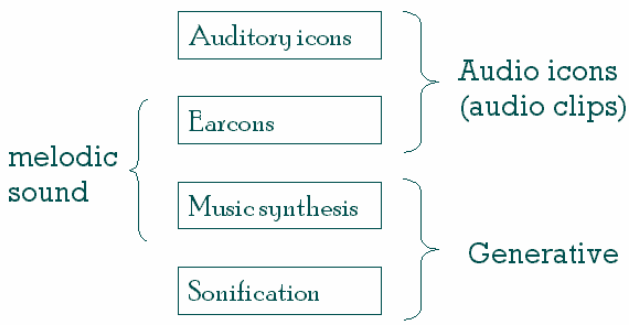

Figure 3. Four types of nonspeech audio

3.1. Concepts and modalities

Allocating content to multiple modalities requires the understanding of the relationship between concepts and modalities, i.e. knowing which modality is most suitable for what concepts. Figure 3 lists the top concepts in the EuroWordNet project [4]. They are divided into three types:

- 1stOrderEntities correspond to concrete, observable physical objects, persons, animals and physical substances. They can be located at any point in time and in a 3D space.
- 2ndOrderEntities are processes, states, situations and events that can be located in time. Whereas 1stOrderEntities *exist* in time and space 2ndOrderEntities *occur* or *take place*, rather than exist.
- 3rdOrderEntities are mental entities such as ideas, concepts and thoughts that exist outside space/time dimension and are unobservable. They can be predicated as true or false rather than real, they can be asserted or denied, remembered or forgotten.

Figure 4. Top Concepts in EuroWordNet

Consider from the prospect of multimodal presentation, 1stOrderEntities are suitable to be presented in static visual modalities (still pictures), 2ndOrderEntities are suitable to be displayed in dynamic visual modalities (animation or video, accompanied with nonspeech audio as a supplement), and 3rdOrderEntities are suitable to be expressed in language (text/speech) since they are unobservable by visual sensor. This classification has a mapping to the linguistic concept *part-of-speech*: the 1stOrderEntities relate with concrete nouns; *static situations* in the 2ndOrderEntities concern either properties of entities or relations between entities in a 3D space, i.e. adjectives and prepositions; *dynamic situations* in the 2ndOrderEntities cover either events or their action manners, i.e. verbs; and the 3rdOrderEntities are covered as *non-action verbs* (e.g. “decide”, “believe”, “doubt”). This paper focuses on the 2ndOrderEntitis (i.e. event verbs and adjectives) and their relations to auditory display.

3.2. Verb ontology and audio semantics

Various English verb classifications have been analyzed in terms of their syntactic and semantic properties, and conceptual components, such as syntactic valency, lexical semantics, semantic/syntactic correlations [4], and visual semantics [5]. Here the audio semantics of verbs, particularly their sources, is studied.

The verb ontology shown in Figure 4 represents a classification of sound emission verbs based on audio semantics. First we divide sound emission verbs into three classes: 1) sounds made by one object, 2) audio expressions of human, and 3) verbs of impact by contact, i.e. sounds made by two objects, based on sound source. In the first class, we classify the verbs to those emitting typical sounds of a particular object (class 1.1), sounds made by animals (class 1.2), those emitting break/split sounds (class 1.3), and weather verbs which emit environmental sounds (class 1.4). Class 2 includes sounds made by human, either speech (class 2.1) or nonspeech expressions (class 2.2). Nonspeech expressions are composed of nonverbal expressions such as “laugh”, “sign” (class 2.2.1), musical expressions such as “hum”, “sing” (class 2.2.2), auditory gestures such as “clap”, “snap” (class 2.2.3), and hiccup/breathe verbs such as “fart”, “sneeze” (class 2.2.4). Class 3, the verbs of impact, includes *nonagentive verbs of impact* (class 3.1), e.g. my car *bumped* into the tree, *contact of an instrument and an object* (class 3.2), and *contact of body part and an object* (class 3.3).

1. Sounds made by one object
  1.1 Typical sounds of a particular (object) source, e.g. toll, (gun) fire, (clock) tick, trumpet
  1.2 Sounds made by animals, e.g. baa, bark, quack, tweet
  1.3 Break/split verbs, e.g. break, crack, snap, tear
  1.4 Weather verbs, e.g. storm, thunder
2. Audio expressions of human
  2.1 Verbs of speaking or manner of speaking, e.g. say, order, jabber, shout
  2.2 Nonspeech expressions
    2.2.1 Nonverbal expressions, e.g. laugh, giggle, moan, sign
    2.2.2 Musical expressions, e.g. hum, sing, play (musical instruments)
    2.2.3 Gestures, e.g. clap, snap
    2.2.4 Hiccup/breathe verbs, e.g. fart, hiccup, sneeze, cough
3. Verbs of impact by contact
  3.1 Nonagentive verbs of impact (by contact of two objects), e.g. bump, crash, slam, thud
  3.2 Contact of one instrument and one object, e.g. strike (with a stick)
  3.3 Contact of body part and one object, e.g. kick, knock, scratch, tap

Figure 5. Verb ontology on audio semantics

The importance of the audio modality varies from class to class. For instance, audio is indispensable for the class 2.1 (Speaking or manner of speaking) and class 2.2.2 (Musical expressions), whereas it is merely an addition to the visual modality for the class 3 (verbs of impact by contact). This information is used in media allocation and animation generation in CONFUCIUS, for example, verbs of speaking or manner of speaking (class 2.1) cause the part enclosed in quotation marks in a sentence being transferred to text-to-speech engine and the simultaneous lip movements of the speaker in generated 3D animation.

3.3. Attribute ontology and audio semantics

Conventional classification of adjectives [6] divides them into two major classes: descriptive adjectives and relational adjectives. Descriptive adjectives (such as *large/small*,

interesting/boring) ascribe to their head nouns values of bipolar attributes and consequently are organized in terms of binary oppositions (antonymy) and similarity of meaning (synonymy). Relational adjectives (such as *nuclear* and *royal*) are assumed to be stylistic variants of modifying nouns and can be cross-referenced to the nouns.

In Figure 5 we classify the category of adjectives according to the perceiving senses they require. The first level is distinguished by the standard whether they can be perceived through visual sense as vision is a main input modality of human perception. Visually observable adjectives are adjectives whose meaning could be directly observed by human eyes. They consist of adjectives describing objects' attributes or states, e.g. dark/light, large/small, white/black (and other color adjectives), long/short, new/old, high/low, full/empty, open/closed, observable human attributes, and relational adjectives. Observable human attributes includes human emotions, such as *happy/sad*, *angry*, *excited*, *surprised*, *terrified*, and other non-emotional features such as *old/young*, *beautiful/ugly*, *strong/weak*, *poor/rich*, *fat/thin*. Human feelings are usually expressed by facial expression and body posture, while non-emotional features are represented by some body features or costumes. This convention is also used in performance art.

Figure 6. Categories of adjectives

The third kind of *visually observed adjectives* is a large and open class--*relational adjectives*. They usually mean something like "of, relating/pertaining to, or associated with" some noun instead of relating to some attribute, and play a role similar to that of a modifying noun. For example, *nasal*, as in *a nasal voice* relates to *nose*, *mural*, as in *mural painting*, relates to *wall*, and *royal* relates to *king* or *queen*. The relational adjective and its related noun refer to the same concept, but they differ morphologically. Moreover, relational adjectives have features like nouns and unlike descriptive adjectives: they do not refer to a property of their head nouns; they are not gradable; they do not have antonyms; and the most important, their visual semantics are the same as their corresponding nouns. Therefore CONFUCIUS treats this subcategory of adjectives as noun, and represents the appropriate nouns that they point to.

There are three types in the unobservable class: the first type is adjectives that can be perceived by other modalities such as haptics, e.g. wet/dry, warm/cold, coarse/smooth, hard/soft, heavy/light, or auditory display, e.g. quite/noisy, loud/soft, cacophonous/euphonious. The second class of visually unobservable adjectives is abstract attributes, either unobservable human attributes concerning virtue (e.g. good/evil, kind, mean, ambitious) or non-human attributes (e.g. easy/difficult, real, important, particular, right/wrong, early/late); the last type is the closed class of *reference-modifying adjectives.* They are a relatively small number of adjectives including *former*, *last*, and *present* etc.

CONFUCIUS represents unobservable adjectives in language and audio modalities. Here we shall distinguish narrator's language with character's language. If the adjective appears in a character's dialogue it is just transmitted to the text-to-speech engine directly and is presented in speech modality; if it appears in the narration part, the natural language processing component judges whether it is presentable visually, and if not, the sentence is sent back to re-allocate to the audio modality (speech or nonspeech sound). The unobservable adjective may be presented by the narrator's voiceover (speech) or nonspeech sounds such as auditory icons, nonverbal expressions, or music.

3.3.1. Entity properties for visual and audio display

WordNet [8] explicitly encodes ascriptive adjectives by describing the attribute to which the adjective ascribes a value. For example, the attribute for "loud" is "volumn". The complete list of attributes can be had by running every adjective in the WordNet adjective index asking for attributes. The result is a list of about 160 unique nouns (synsets) that are used as attributes. Based on [8] and [9], we summarize the following list of "visually representable" and "audio representatble" properties in Table 1 and 2 respectively. The semantic analyser of CONFUCIUS tags these properties of ascriptive adjectives in natural language processing.

Classes of properties	Properties
Space	size, length, width, thickness, height, depth, orientation (direction), shape (form), texture, speed (rate)
Matter	density, state (of matter), appearance, color, quantity (numerousness)
Time	Timing
Human attributes	gender, age
Affection	affection, sensitivity, emotion, personality, quality

Table 1: Visually representable properties

Table 2 lists the audio representable properties and types of auditory display which can be used to present these properties. Some of them are straightforward, the sound dimension properties can be presented by all types of auditory display, for instance. Audio presentation of some properties is indirect, especially the matter properties and affections. For example, brittleness could be displayed through auditory icons of break/split sounds, and emotions can be conveyed through music.

4. AUDIO IN CONFUCIUS

CONFUCIUS' audio presentation includes auditory icons and text-to-speech. Auditory icons are encapsulated in the 3D models of objects and virtual humans, e.g. the firing sound of a gun is encapsulated in the gun's geometric file, and the hiccup and yawn sounds of a virtual character are encapsulated in his/her VRML file. These auditory icons accompany animated events in the story being told. Since auditory information can be redundant with visual and language modalities, determining whether to eliminate the visual (or speech) information or make

the audio information redundant is a task of the media allocation module in CONFUCIUS.

Classes of properties	Properties	Types of audio
Sound dimension	frequency (pitch), amplitude (volume), timbre	Auditory icons, speech, music
Spatial relation	position, orientation (direction), speed (rate)	Auditory icons, speech, music
Time	duration, timing	Auditory icons, speech, music
Matter	density, state (of matter), quantity (numerousness), weight, brittleness	Auditory icons
Human attributes	gender, age	Speech, nonverbal expressions, music expressions
Affections	affection, sensitivity, emotion, personality, quality	Speech, music

Table 2. Audio representable properties

4.1. Media allocation

Multimedia integration requires the selection and coordination of multiple media and modalities. The selection rules are generalized to take into account the system's communicative goal, features characterizing the information to be displayed and features characterizing the media available to the system. To tell a story by complementary multi-modalities available to CONFUCIUS the system concerns dividing information and assigning primitives to different modalities according to their features and cognitive economy. Since each medium can perform various communicative functions, designing a multimedia presentation requires determination of what information is conveyed by which medium at first, i.e. allocating contents to media according to *media preferences*. For example, presenting spatial information like position, orientation, composition and physical attributes like size, shape, color by visual modality; presenting events and actions by animation; presenting dialogue/monologue and temporal information like "ten years later" by speech; presenting dog bark by both audio and visual modalities (or by audio icon solely). We formulate the principles for media allocation within CONFUCIUS as the following:

1. Realize spatial information, physical attributes, physical actions and events in 3D animation.
2. Realize dialogues, monologues, and abstract concepts (including abstract actions and abstract relations) in speech (including voiceover narrative). For example, if the media allocator detects an unobservable adjective that can not be presented visually in the narration part of a story, the sentence is sent to the presentation agent and is output in Merlin the narrator's speech and gestures, while the other visually presentable parts of the sentence are still allocated to the animation engine and TTS to create animations which will be played when Merlin is talking.
3. Realize (or augment other modalities) sound emission verbs in Figure 5 and audio representable adjectives in Figure 6 in audio modality.
4. Realize failed attempts (e.g. animation files not available) and successful attempts with low confidence in principle 1 in other modalities according to the feedback from the animation engine.

Figure 7 shows CONFUCIUS' multimedia presentation planning. Media allocation also receives feedback from media realization, such as animation engine, to influence the selection of media for a definite content. Thus we allow a decision made or failed realization at a later stage of processing can propagate back to undo an earlier decision. For example, realization in animation engine may fail because of visualisation difficulties, and this message should be fed back to *media allocator*, where the content could be re-allocated to other media. Media coordination includes cross modal reference (e.g. Merlin refers to the virtual humans in animation), synchronicity (e.g. lip-speech synchronising), and timing (e.g. scheduling Merlin and virtual humans' speech/movements). Finally, media player consists of VRML player and Javascript for Microsoft Agent control. The former is for playing virtual characters' animation and speech, and the latter is for playing Merlin's speech and behaviours.

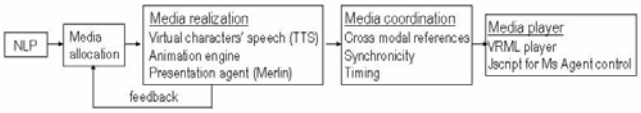

Figure 7: CONFUCIUS' multimedia planning

4.2. Text-to-speech

There are two ways to synchronize a character's lip animation with his speech (through a TTS engine). The first is to obtain estimates of word and phoneme timings and construct an animation schedule prior to execution (time-driven). The other is to assume the availability of real-time events from the TTS engine-generated while the TTS is producing audio, and compile a series of event-triggered rules to govern the generation of the animation (event-driven). The first approach allows us to choose a TTS engine more freely, while the second must be used with TTS engines supporting event-driven timing, such as Microsoft Whistler [10].

We use FreeTTS for speech synthesis because it is written entirely in the Java programming language, supports Java Speech API, and fits well to our developing environment. FreeTTS is derived from the Festival Speech Synthesis System from the University of Edinburgh and the FestVox project from Carnegie Mellon University. The algorithm of the program interfacing with FreeTTS is described as the following:

1. Find a pair of quotation marks
2. In the context, looking for a verb in the class 2.1 in Figure 5 or a verb in the WordNet group <verb.cognition>

Covering cognition verbs ensures that the speech modality takes charge of what a character is thinking. It is common in temporal visual arts like movie or cartoon that a character actually speaks out what (s)he is thinking. Here are three examples taken from *Alice in Wonderland*.

Example 1: "You ought to have finished," *said* the King. "When did you begin?" (typical verb of speaking)

Example 2: "I beg pardon, your Majesty," he *began*, "for bringing these in. But I hadn't quite finished my tea when I was sent for." (One sense in WordNet is "begin to speak or say". It belongs to <verb.communication>.)

Example 3: "And that's the jury-box," *thought* Alice, "and those twelve creatures." (<verb.cognition>)

Visual modality is used to differentiate between verbs of speaking and verbs of thinking (<verb.cognition>). Though both contents are expressed by speech, thinking verbs do not accompany with lip movements.

3. Find the speaker, gender, age, give it an ID (name) for a specific voice.

Here the Java Speech API Markup Language (JSML) is used to annotate text input to speech synthesizers. JSML elements provide a speech synthesizer with detailed information on how to speak text and thus enable improvements in the quality, naturalness and understandability of synthesized speech output. JSML defines elements that indicate phrasing, emphasis, pitch and speaking rate, and control other important speech characteristics.

5. CONCLUSIONS

A number of projects are currently based on combining multiple media including animation and speech, exploring a variety of applications in different domains such as intelligent agents [11], virtual theatre [12], virtual human [13], and interactive storytelling [14]. However, few of these systems take the modern NLP approach that an intelligent multimedia system should be based on. CONFUCIUS is an overall framework of intelligent multimedia storytelling, which makes use of state-of-the-art techniques of 3D animation and text-to-speech with the addition of auditory display to achieve realistic virtual storytelling.

We have investigated current nonspeech audio display and describe the use of speech and nonspeech sound in the virtual storytelling of CONFUCIUS, which converts natural language sentences to a virtual story world combining 3D animation, speech and nonspeech audio. A linguistically-based approach concerning lexical semantics of sound emission verbs and audio representable adjectives was introduced. We have discussed several issues such as relations between concepts and multiple modalities, verb ontology and audio semantics, and attribute ontology. The contribution of CONFUCIUS lies in generation and combination of 3D animation, speech and nonspeech sounds from natural language by automating the processes of language parsing, semantic representation, media allocation and realisation. Since this is an ongoing project, future work should include performing test-suite based and user-centered evaluation. We believe that introducing linguistic knowledge has the potential to have an impact on various areas such as intelligent multimedia systems, computer games, movie/ animation production, and virtual environments.

6. REFERENCES

[1] R. Graham, Use of auditory icons as emergency warnings: evaluation within a vehicle collision avoidance application. *Ergonomics*, 42, 1233-1248, 1999.

[2] W. Gaver, Auditory icons: Using sound in computer interfaces. Human Computer Interaction, 2, 167-177, 1986.

[3] M. M. Blattner, D. A. Sumikawa, R. M. Greenberg, Earcons and icons: Their structure and common design principles. *Human-Computer Interaction*, 4, 11-44, 1989.

[4] P. Vossen, L. Bloksma, H. Rodriguez, S. Climent, N. Calzolari, A. Roventini, F. Bertagna, A. Alonge and W. Peters, The EuroWordNet Base Concepts and Top Ontology. EuroWordNet LE2-4003, Deliverable D017, D034, D036, WP5, http://www.illc.uva.nl/EuroWordNet/corebcs/topont.html, 1998.

[5] B. Levin, *English verb classes and alternations: a preliminary investigation*. Chicago: The University of Chicago Press, 1993.

[6] M. Ma and P. McKevitt, Visual semantics and ontology of eventive verbs. Natural Language Processing - IJCNLP-04, First International Joint Conference, Hainan Island, China, March 22-24, 2004, Keh-Yih Su, Jun-Ichi Tsujii, Jong-Hyeok Lee and Oi Yee Kwong (Eds.), 187-196, *Lecture Notes in Artificial Intelligence* (LNAI) series, LNCS 3248. Berlin, Germany: Springer Verlag.

[7] D. Gross and K. Miller, Adjectives in WordNet. *International Journal of Lexicography* 3(4): 265-277, 1990.

[8] C. Fellbaum (Ed.), WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press, 1998.

[9] K. Barker, Object Properties inspired by various resources, http://www.cs.utexas.edu/~kbarker/working_notes/properties.html, 2004.

[10] X. Huang, A. Acero, J. Adcock, H. W. Hon, J. Goldsmith, J. Liu, and M. Plumpe, Whistler: A trainable Text-to-Speech system. Proceedings 4th International Conference on Spoken Language Processing (ICSLP '96), Piscataway, NJ, 2387-2390, 1996.

[11] A. Nijholt. Issues in multimodal nonverbal communication and emotion in embodied (conversational) agents. In: Proc. 6th World Multiconference on Systemics, Cybernetics and Informatics. Volume II: Concepts and Applications of Systemics, Cybernetics and Informatics I. N. Callaos, A. Breda & Y. Fernandez (eds.), International Institute of Informatics and Systemics, July 2002, Orlando, USA, 208-215.

[12] K. Perlin and A. Goldberg. Improv: a system for scripting interactive actors in virtual worlds. In *SIGRAPH'96 Conference Proceeding*, 205-216. 1996.

[13] N. Badler, C. Erignac, and Y. Liu. "Virtual humans for validating maintenance procedures, Comm. of the ACM, Vol. 45, Issue 7, Pg. 56-63, July 2002.

[14] A. Bobick, S. Intille, J. Davis, F. Baird, C. Pinhanez, L. Campbell, Y. Ivanov, A. Schtte, A. Wilson, The KidsRoom: A perceptually-based interactive and immersive story environment, PRESENCE: Teleoperators and Virtual Evironments, 8(4), August 1999, 367- 391.