CSLML: a markup language for expressive Chinese sign language synthesis

By Kejia Ye, Baocai Yin and Lichun Wang*

This paper presents a Chinese Sign Language Markup Language (CSLML), which is developed for expressive Chinese sign language synthesis by introducing features and structure of sign language prosody. The tags of CSLML are divided into two levels: function level and phonetic level. Function level provides abstract information about signed content and prosody, so it facilitates text annotating for text-driven automatic synthetic system and adapts to diversified synthetic methods, such as motion capture animation or image-based synthesis, which may not be good at processing lower-level information. Phonetic level provides detailed behavioral manners based on phonetics and phonology to interpret the meaning in function level. It facilitates creation and edit of any motions. The two levels co-exist in CSLML documents and high-level description can be mapped into corresponding low-level behavior to provide one-to-many variability and expression of synthesis. Therefore, we also introduce a framework for this mapping processing and exhibit results of our animated prototype system based on this framework. Copyright © 2009 John Wiley & Sons, Ltd.

Received: 25 March 2009; Accepted: 26 March 2009

KEY WORDS: Chinese sign language synthesis; representation language; virtual human

Introduction

Overview

Sign language is one of the natural human languages, a visual language, mainly depends on manual gestures. However, manual signs cannot convey all messages contained in spoken language, so it still needs other parts of body such as face, head, and torso to work as assistant articulators while “speaking.” With the cooperation of multi-modal behavior, sign languages express both grammatical and semantic content and other less linguistic information, such as emotion and personal style. Moreover, multi-modal co-articulations contain abundant prosody,1,2 which is the key factor in acceptance and realism of synthetic sign languages. Therefore, information included in representation of sign language can be divided into content information and prosody information. The content information specifies a sequence of non-verbal behavior to construct a discourse, according to sign language grammar (phonology, morphology, lexicon, syntax, etc). It’s the foundation and invariant component of sign language expression. Prosody, the variant component, is influenced by both internal and external linguistic factors. External factors add emotion, register, and personal features to the content, while internal factors add phonological variation in how the sign units are produced to form a continuous and effective stream. Besides, according to articulators, description of sign language should cover manual information and non-manual information. In sign languages, both content information and prosody information are delivered through co-expression of these two aspects.

In existing sign language synthesis, there are three main stages in the processing pipeline: First, natural language processing, translating the spoken language sequences into sign language structure; second, sign language behavior planning, deciding what behavior...
should be exhibited and how they are synchronized; third, sign language behavior synthesis, displaying continuous signs to users, either by virtual human animation or by video images. There are two interfaces between the adjacent phases, phonetic-level (or parameter-level) interface and rendering-level interface. The term “phonetic” here refers to the phonemes (parameters), the smallest units of language. It concerns how the units or phonemes are organized in a language, how they are combined to form a word, and how these phonemes interact with each other. The equivalent forms in sign language are five phonemes identified by William Stokoe: hand shape, palm orientation, movement, location, and non-manual markers. Linguists have developed some effective notation system to record and describe the parameter information of sign language. Originally, it’s just intended to provide a complete description of sign language in written form; however, these notation text have been applied as phonetic-level interface between the first two stages in sign language synthesis pipeline or functioned as a driving script. Since the intuitive symbolic pattern is hard to be processed by computers, most of today’s sign language synthetic systems design a XML-based language to function as an equivalent to the corresponding notation. These XML-based description languages not only contain the five phonemes which are the core components in linguists’ notation, and also supplement some instructive information for sign synthesis.

However, existing sign language annotations or markup languages lack effective description on prosodic features and meaningful non-manual behaviors of sign language, they usually functioned in content information, especially at phonetic level, which directly provide such detail of multi-modal behavior as the four manual phonemes or specific movements of other parts of the body. Since sentences with the same text can be expressed with different emotions and personal feelings, which means the prosodic behavior may be diversified based on the same sentences, the way of directly defining phonetic behavior may limit the variability and acceptance of synthesis. For example, sometimes signers may focus on some words of a sentence to express a special semantic meaning. Different places where they put the emphasis on may result in different effects to the listeners. Especially, prosody in visual language is conveyed through very complex fashion and the deaf person’s signing looks like a mime which is a smart cooperation of the whole body, so they can express question utterance through tilted head, raised brows, or something else they choose to do based on context or personal styles. Besides the lack of flexibility and expressive capability, the phonetic-level scripts also required intensive labor work and abundant knowledge of annotation and it covers complex combination of multiple channels, so it may have difficulty in generating the script automatically. Hence, higher level description is needed to extend the expression of script and to provide flexibility through diverse mapping mechanism. What’s more, there has been lots of successful experience of translating plain-text into functional, abstract, and high-level markup language in TTS and ECAs system to support the automatic synthesis system.

At present, the notation system of Chinese Sign Language (CSL) in linguistics hasn’t been fully developed. Since the existing notations may not exactly express some characteristics in CSL such as the Character Signs or Writing-In-The-Air, a lot of effort to develop an effective and integrated notation of CSL remain under progress. Hence, according to special situation of CSL development and limitations of existing markup language, we present a new representation language for the Chinese sign language synthesis—Chinese sign language markup language (CSLML). CSLML pays more attention on the prosody information and introduces an abstract function level, which is independent form realization and focuses on high-level content and prosodic information, into the traditional phonetic (parameter) level marks. Our goals are: (1) to build up a foundation and general interface for follow-up researches of CSL synthesis; (2) to provide a concise, complete, and effective CSL description for expressive synthesis; (3) to improve the variability of synthesis and make it suitable for different kinds of realization (animation or video images) by defining a two-level script and an applied framework; (4) to provide a capability of creating scripts from manual editor or automatic generating from textual input.

Related Work

At present, most annotation systems and their markup languages focus on phonology, especially manual gestures. Hamburg Notation System (HamNoSys) and the Signing Gesture Markup Language (SiGML) is a representative pair, SignWriting and SignWriting Markup Language (SWML) is another pair. HamNoSys, a well-established notation originally developed by the University of Hamburg for the transcription of sign languages at the phonetic level. Manual phonemes are described as hand-shape, hand orientation, hand position and motion. Its character repertoire contains over 200 iconic symbols, for which a computer font is available. Based on the
phonetic model, SiGML is a XML-compatible transliteration of HamNoSys itself. SiGML organizes descriptive sequences in a manner suitable for generation of animated avatar and provides flexible creation and edit of manual gestures. However, SiGML only concerns the generation of phonetic-level aspects of signing, both manual and non-manual, other aspects of sign language are still missing, such as prosody and some expressive behavior, which may not be strictly linguistic but play an important communicative role. SignWriting is a writing system which uses visual symbols to represent handshapes, movements, and facial expressions of sign languages. It is an “alphabet”—a list of symbols used to write signed language. SWML is an XML-based format, developed for the storage and processing of SignWriting texts and dictionaries. It allows inter-operability of SignWriting savvy applications and promotes web accessibility for deaf people in their own natural languages. However, SignWriting also lacks a prosody description and its dictionary has been mainly focused on Western sign languages. Besides, there are many other sign language annotation systems which are designed for specific language or specific systems. For instance, The GesSyCa system\(^9\) for French Sign Language uses QualGest notation system. Most of these XML-based markup languages just work as a translation or Unicode form of corresponding annotation systems, so they all lack the prosodic features and synchronization mechanism for multi-modal behavior. Furthermore, they are fully developed for a limited number of sign languages, such as ASL and BSL, and may be invalid or inaccurate on some special phenomena of CSL. Thus the researches of CSL really need an annotation system and an equivalent markup language for automatic sign synthesis.

The expression of prosody and emotion, a lack in the sign language synthesis, is always an emphasis of Embodied Conversational Agents (ECAs) researches. In ECA, multi-modal non-verbal behavior assisted in verbal speech function as random variables to enhance the expression and realism of avatars. It belongs to the second information recorded in sign language synthesis, non-grammatical information, the prosody. Now ECA systems separate its description languages into two categories. Abstract level focuses on description of speaker’s intent, emotion, and personality; whereas the detailed cooperation of behavior is recorded by another low level language. This separation ensures the independence between mind module of agent and animation player. Affecitive Presentation Markup Language (APML)\(^{10}\) is a typical high-level description which specifies an agent’s behavior at the meaning level. It’s based on Poggi’s taxonomy of communicative functions which are defined as a meaning-signal pair. A communicative function may be associated with different signals. That is for a given meaning, there may be several ways to communicate it. Another type of markup in ECA is designed to specify communicative and expressive behaviors. Behavior Markup Language (BML)\(^{11}\) works as an interface between modules of behavior planning and realization, providing general, player-independent description of multi-modal behavior, staying above specific implementations. It provides sufficient details in describing behavior, from occurrences and relative timing of involved actions, to detailed definitions of behavior’s form. Moreover, it still has flexibility for future expanse. Other examples of such languages are RRL\(^{12}\) and MURML\(^{13}\). They are all designed for specific system and cannot be regarded as a universal description. It’s clear that the strong ability to generate expressive behavior in ECA is worth to be referenced in sign language synthesis. However, ECA pays more attention on providing information for autonomous planning of agents’ behavior. Although ECA can integrate some sign language notation, it may still lack enough supports for special grammatical phenomena of sign language, and a large number of information is redundant for sign synthesis.

Accordingly, in this paper we combine advantages of existing description methods in ECA and sign language annotation and propose a new markup language to addresses the issue of lacking a general, effective, and concise annotation for CSL synthesis, which is more expressive in describing prosodic features and cooperation of multi-modal behavior, and is more suitable for grammatical structure and characteristics in CSL.

**CSLML Specification**

In this section we present a markup language for expressive CSL synthesis—Chinese sign language markup language (CSLML).

**Features Overview**

**Two-level Structure.** CSLML is designed as two levels: function level and phonetic level. Function level, the top level of CSLML structure, includes abstract information about content descriptions and prosodic marks. Phonetic level, below the function level, interprets how to “speak out” the function-level information through co-articulations of multiple articulators in CSL. All the
CSLML tags can be categorized into these two levels and all tags can coexist in one document. The CSL synthetic architecture presented in next section will explain how to use the CSLML at different level to increase the diversity of synthetic signing behaviors and the flexibility of CSLML document generation.

Two Category Content. The tags in CSLML, no matter in which level, can be categorized into two classes: content mark and prosodic mark. Content mark includes all behavioral description defined in the CSL grammar standard, which shows the correct manner accepted by most Chinese deaf people to sign a word in CSL. It consists of the phonemes of basic units (roots) and the three-layer structure in CSL (root, word, and sentence). It only provides the name or index of the basic sign at function level and leave the detail of phonemes to the phonetic level. Prosodic marks consists of four types of parameters at function level: stress, utterance, emotion, and phonological_process. Stress is usually used to express the semantic center or key information of a sentence. Utterance is mainly expressed by non-manual behavior, especially facial expressions. Emotion is a short-time mood related to context and personality. Especially, the parameter phonological_process referring to special phonological phenomena (such as movement reduction or anticipation) which plays an important role in generating natural prosody, is left for future supplement. Besides, according to Chinese prosody hierarchy studied in Chinese TTS, prosodic phrases also exist in CSL. Hence, we add a level prosodic_word into the three-level structure tags. The four-level structure defines the break pattern between phrases, words, or sentences to show the natural rhythm or subconscious turn-taking. All these function-level prosodic marks aim to improve the acceptance and realism of synthetic system. Moreover, prosodic marks at phonetic level are defined through multi-modal non-manual behavior and manual prosodic parameters. The

Figure 1. The organizational structure of main tags in CSLML.
detailed elements are shown in Figure 1. Since the variability is mainly related to prosody, the more we can realize through combination of phonetic-level elements, the more expressive and acceptable synthesis we can get.

**Definition of Tags**

```
<!ELEMENT CSLML (sign_group+ | basic_sign+))
<!ELEMENT sign_group(#PCDATA | sign_group | basic_sign | prosody | prosody_nonmanual)+
<!ATTLIST sign_group id ID #IMPLIED type (sentence | prosodic_word | word) #REQUIRED
<!ELEMENT prosody_manual (duration, hold, repetition?, variation?)
<!ELEMENT prosody_nonmanual (sign_nonmanual)
<!ELEMENT BASIC SIGN. ——)
<!ELEMENT basic_sign (prosody?, ((sign_manual, sign_nonmanual?), (prosody_manual?,))
<!ATTLIST basic_sign id ID #IMPLIED gloss %basic_sign_namespace;
<!ELEMENT sign_manual (manual_phonomes?, prosody_manual?)
<!ATTLIST sign_manual gloss %sign_manual_attributes;
<!ELEMENT sign_nonmanual (head | face | lips | body | eyegaze | extra+)
<!ELEMENT head (head_pos | head_move)+
<!ELEMENT head_move EMPTY
<!ATTLIST head_move id ID #IMPLIED type %head_movement; amount CDATA “0” start CDATA #IMPLIED end CDATA #IMPLIED syn %synchronization;
<!ENTITY %basic_sign_namespace (neutral | TA | Tongguo | Y1 | ...) “neutral”)
```

**Structure Format Tags.** Every CSLML script should start with the root tag (*cslml*). The following tags can only be either (*sign_group*) or (*basic_sign*). A (*basic_sign*) represents a CSL root, basic unit of CSL. A (*sign_group*) consists of a set of (*basic_sign*) in the signed order. This tag is used to work as a bound which lets the related roots share the same attributes labeled by the attribute *type*. The *type* can only take three values: “word,” “prosodic_word,” and “sentence,” which indicates the prosody hierarchy. For example, roots in the same word may have same stress intensity, roots in same sentence may share the same emotion.

**Prosody Tags.** There are two levels in prosody. (*prosody*) provides information at function level, while (*prosody_manual*) and (*prosody_nonmanual*) work at lower level. (*prosody_manual*) defines specific prosodic parameter related to manual movements, consisting four prosodic parameters: duration, hold, repetition, variation. The element variation is a extensible interface to accept other special changes of manual movements. Besides, (*prosody_nonmanual*) defines detail of cooperation among multi-modal behavior. All prosody tags can appear as sub elements of both (*sign_group*) and (*basic_sign*) or even the lower level of basic_sign to change specific segments of signs, but they may result in some conflicts sometimes. There is no universal resolution for all the different conditions; however, the regular rule is that lower level in prosodic hierarchy takes precedence over the higher level or blend the two effects to form a integrated prosody.

**Basic Sign Tags.** A basic sign consists of content information and prosody information (Figure 1). Content information is composed of manual gestures and non-manual postures, specified with (*sign_manual*) and (*sign_nonmanual*), respectively. The attribute *gloss* provides a name or index of the root, which is a function-level information of manual gesture. The entity (*basic_sign_namespace*) contains all names of CSL roots.

**Manual Sign Tags.** (*sign_manual*) is used to provide parameter information about a specific root. In China, the phonology study has not been mature and some typical features in CSL still cannot be displayed intuitively by existing notation systems. Hence, this part in CSLML we set aside as an interface to be expanded in the future. At present our prototype system and most CSL synthetic system just use the function level information or root indexes to finish the signing synthesis.
Non-manual Sign Tags. \((\text{sign\_nonmanual})\) tag defines how different articulators in CSL cooperate with each other. In CSL, there are five main non-manual articulators: head, face, lips, body, and eye gaze. The tag \((\text{sign\_nonmanual})\) contained in \((\text{basic\_sign})\) try to describe behavior specified by CSL grammar standard. However, tags in \((\text{prosody\_nonmanual})\) represents those behavior caused by prosodic requires, such as emphasis, affective feelings, and so on. Each articulator may consist of several parts (Figure 1). We can define a facial expression through general description like glad or separately define the movement of each organ. In CSL, eye brows, eye lids, and mouth are the main articulators which play an important role in delivering the meaning and emotion. Besides, movements of body are largely driven by shoulder and torso, which usually accompany the pace of signing and reveal the characteristics of a signer. We have given a definition of head in the DTD above and we will exhibit the practical usage through an example of CSLML segment in next section.

**Architecture of Synthesis Driven by CSLML**

**Generation and Application**

An original CSLML script may be created manually through a CSLML supported editor tool or translated automatically from Chinese input through natural language processing and semantic analysis. The creators, both human and machine, may generate CSLML documents marked-up with tags of different levels. In the worst situation, the generated documents only provide content information at function level (index of basic sign). Adversely in the best situation, creator may skip the function-level markup (both content and prosodic) and produce phonetic-level markup for partial or entire documents. For instance, some “copy synthesis” or “prosody transplant” tools, which copying properties from videos to animation, will require the full-detailed mark-up. We already discussed that the more detailed prosodic information added into synthesis, the more expressive and natural synthesis will be achieved. Hence, the CSL synthetic architecture shown in Figure 2 represents a general framework about how to use the CSLML at different level to increase the diversity of synthetic signing behaviors.

When a synthesis module parses the CSLML, the information contained in function-level tags can be recognized. If CSL synthesizing module owns a phonology planning sub module, which has the ability to transform the abstract information at function level into detailed phonetic-level behavior, it can add specific description of prosodic behavior into the script, which may be very useful to animated system. Otherwise, the high-level tags may be directly ignored by some system without the phonology planning capability or just be directly processed according to system needs, such as some motion capture or key-frame animation system. In the framework, the flexible generation of prosodic behavior is dependent on a database in the CSL phonology planning module. This database stores the CSL prosody rules and CSL grammar details, which decide how to map a prosodic mark at function level into a set of multimodal behavior. For example, in CSL the utterance of question cannot be conveyed solely by manual gestures, but expressed through facial expressions and body gestures. The typical features to show a yes–no question are raised brows, forward-tilted head. Hence, the database may contain a piece of rule to map the utterance of yes–no question into the corresponding facial and body behav-

![Figure 2. The transition framework from function-level CSLML to phonetic-level CSLML during CSL synthetic processing.](image)
ior. However, mapping rules are one-to-many, a powerful mapping algorithm will greatly improve expression of synthetic results.

**Examples**

In this section, we give an example about how to use CSLML to express the content expected to be synthesized. Following CSLML segment describes the sentence “她们真漂亮！(They are so beautiful)” in CSL. The adverb “真(so)” is replaced by intensive behavior while signing the related adjective, so only two CSL words need to be signed “她们(they)” and “漂亮(beautiful).” “漂亮(beautiful)” is a one root word and the root is named by “Mei.” “她们(they)” consists of two roots “Ta” and “Dajia.” Therefore, we use ⟨sign group⟩ to bound the three ⟨basic sign⟩ together and two ⟨prosody⟩ to indicate high level prosody features, an emphasis at “beautiful” with strong attitude and a joyful feeling accompanying the sentence.

VICE 2009; 20: 237–245

CSLML: A MARKUP LANGUAGE

Firstly, a ⟨prosody manual⟩ tag is added to the basic sign “Ta.” The hold time is reduced to minimum, due to the hold deletion phenomena when a root is combined with other roots to form a larger segment. And the duration value here is a relative amount according to specific system needs. Secondly, a sequence of visemes is added to the first word to accompany the manual signing, as the mouthing of spoken words is important in CSL expression. ⟨sequence⟩ tag indicates that all the ⟨lips⟩ contained in it should be executed orderly. Thirdly, a nod of head is generated when mapping the utterance of intensification. This result can be clearly seen in Figure 3. Besides, a ⟨prosody nonmanual⟩ is added to the end of the sentence. A ⟨face⟩ expression could deliver the emotion “joy” effectively. The attributes start and syn is used to show this movement starts at the point marked with the unique id “s1,” which is the mark indicated the beginning of this sentence. The attribute amount describes the intensity of the movement.
Usage: CSL Synthetic System Driven by CSLML

Based on above work, we have developed a CSLML driven CSL synthetic system: Chinese Sign language Expressive Animated Broadcaster (CSLEAB). CSLEAB takes Chinese text as input and generates raw CSLML documents, which contains content marks at function level (indexes of basic signs). Since we have built up a manual motion database of all CSL roots through data-gloves, we can just use the indexes to splice them together. As manual gestures are necessary components in sign languages and every other behavior always accompany a specific manual segment, root, word, phrase, or sentence. Hence, we firstly generate manual animation parameters along the time-line and then produce parameters of other parts synchronized with the manual start points and durations. The raw markup document may only contain few prosodic information, such as prosodic hierarchy which can be deferred automatically based on the syntax. Therefore, additional prosody has to be added manually through a multi-modal behavior editor tool.

CSLEAB realizes conversion from abstract description to sign parameters based on the framework in last section. Only a few simple rules can be processed right now, like stress to speed and prosody hierarchy to hold level. To get more expressive synthetic signs we need more phonology rules to be learned from real signing corpus.

Conclusions

In this paper, we propose a XML-based CSLML, which is designed towards expressive Chinese sign language synthesis. CSLML contains two types of information: content-related and prosody-related. In particular, we focus on importance of separating abstract description from phonetic elements, discuss interface between the two levels, and provide a general framework for transformation from function level to phonetic level. Finally, we exhibit our animated prototype system CSLEAB driven by CSLML scripts. Although the description and its applied framework have provided the capability to synthesize much more expressive and natural Chinese sign language than ever before, a lot of effort remain required to really demonstrate this effect. Developing a more effective and efficient algorithm for automatic generation of CSLML and conversion between the two levels is key to the issue. Besides, construction of more accurate and expressive motion data is another important aspect to support the descriptive mechanism. However, the CSLML has not been fully finished and it will keep developing as the Chinese sign language research progresses.

ACKNOWLEDGEMENTS

This work was supported by BJUT Multimedia and Intelligent Software Lab, NSFC (60533030, 60825203), and Scientific Research Common Program of Beijing Municipal Commission of Education (KM200710005023).

References


Authors’ biographies:

Kejia Ye is a master student of Beijing Key Lab of Multimedia and Intelligent Software Technology, College of Computer Science, Beijing University of Technology. Her research area is human computer interaction and virtual human animation. (Email: christie.ykj@gmail.com)

Professor Baocai Yin is director of Beijing Key Lab of Multimedia and Intelligent Software Technology, College of Computer Science, Beijing University of Technology. His research interests include digital multimedia, multi-functional perception, virtual reality, and computer graphics. (Email: ybc@bjut.edu.cn)

Lichun Wang is associate professor of Beijing Key Lab of Multimedia and Intelligent Software Technology, College of Computer Science, Beijing University of Technology. Her research interest is human computer interaction. (Email: wanglc@bjut.edu.cn)