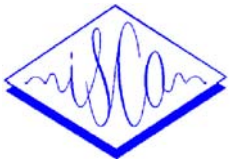

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

Representation and processing of linguistic structures for an all-prosodic synthesis system using XML

Mark Huckvale
University College London
Gower Street, London, U.K.
M.Huckvale@ucl.ac.uk
http://www.phon.ucl.ac.uk

10.21437/Eurospeech.1999-403

ABSTRACT

The ProSynth speech synthesis project aims to re-implement and extend the YorkTalk all-prosodic synthesis system in an open manner preserving its most appealing theoretical aspects. A significant novel aspect of the architecture of ProSynth is the use of the extensible mark-up language (XML) as a computational formalism for the representation of hierarchical linguistic structures. The facilities provided by XML match closely the requirements to represent the phonological features of an utterance in a metrical prosodic structure, namely: nodes described by attribute-value pairs forming strict hierarchies. The XML formalism also leads to an elegant and efficient method for representing declarative phonological contexts under which phonetic interpretation is performed.

1. INTRODUCTION

The ProSynth speech synthesis system [3] is a new implementation of all-prosodic speech synthesis derived from the YorkTalk system [8]. Whereas YorkTalk was implemented in Prolog and was specifically designed for formant synthesis, the ProSynth system has an open and portable architecture compatible with a number of signal generation methods. ProSynth has been used to drive diphone concatenation, a quasi-articulatory model and prosody-manipulated natural speech. Details of its linguistic processing are being reported elsewhere [5][9].

In the new implementation of YorkTalk we have sought to preserve its most theoretically appealing characteristics: its non-linear framework for data representation and its declarative formulation of the knowledge required for phonetic interpretation of phonological structures. Phonological representations in YorkTalk took the form of a hierarchical metrical structure of utterance, foot, syllable, onset, rhyme, nucleus and coda. Features on these nodes were then *interpreted* to create a parameter table for the Klatt formant synthesizer without reference to the phoneme labels at the bottom of the hierarchy. Phonetic interpretation took the form of declarative statements that expressed relationships between the context in which a phonological component was positioned and its consequent acoustic form or *exponency*. Phonetic interpretation had two clear phases: *temporal interpretation* where the durations and the overlapping of phonological components were established, and *parametric interpretation* where the acoustic-phonetic forms of the components were realised according to their context and duration. The overall result was speech synthesis of single foot utterances which had remarkably natural rhythm.

On the other hand, YorkTalk had a number of limitations. It was implemented in Prolog and had grown to a size and complexity that made it hard to develop further. The specific characteristics of the Klatt formant synthesizer were strongly embedded in its parametric interpretation component. It was not a complete text-to-speech system: it merely interpreted marked phonological transcription of single foot utterances. In some ways it could be seen as a 'proof of concept' rather than a general purpose speech synthesis framework.

2. NEW IMPLEMENTATION DESIGN

The aims of the ProSynth project are to build on the knowledge gained in YorkTalk; to create an open computational architecture for synthesis; to extend the system to multiple-foot phrases; to add in an intonational component; to add in capability for longer distance co-articulatory effects; and to support a number of signal generation methods. For the new system, it was considered best to re-use the knowledge gained in building YorkTalk rather than the implementation details.

The new implementation is written in 'C' rather than in Prolog. However there is still a clear separation between the computational engine and the computational representations of data and knowledge. The overall architecture is shown in Figure 1.

Text marked for the type and placement of accents is input to the system, and a pronunciation lexicon is used to construct a strictly layered metrical structure for each intonational phrase in turn. The overall utterance is then represented as a hierarchy as follows:

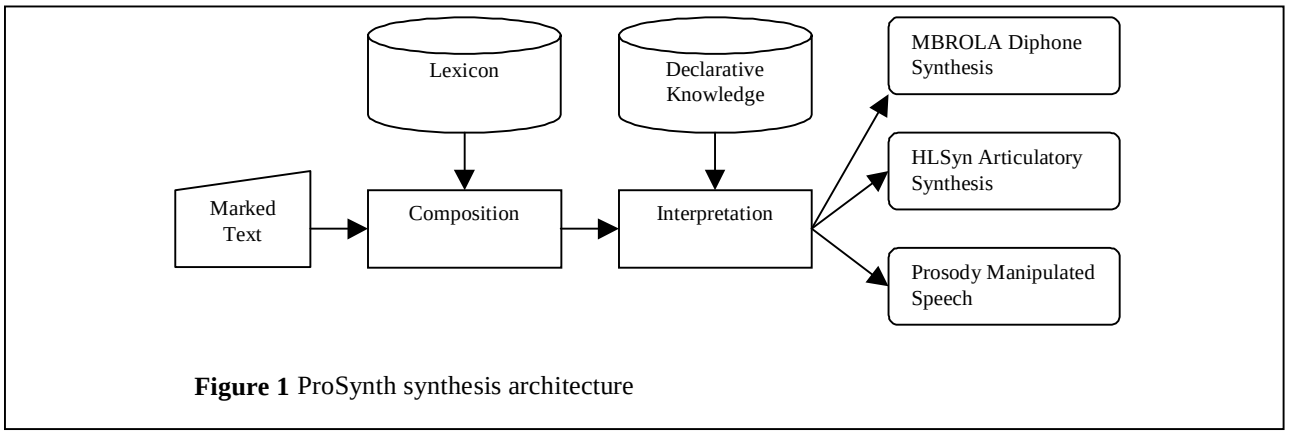

Figure 1 ProSynth synthesis architecture

```
Utterance (UTT)
  Word Sequence (WORDSEQ)
    Word (WORD)
  Intonational Phrase (IP)
    Accent Group (AG)
      Foot (FOOT)
        Syllable (SYL)
          Onset (ONSET)
            Consonant (CNS)
          Rhyme (RHYME)
            Nucleus (NUC)
              Vocalic (VOC)
            Coda (CODA)
              Consonant (CNS)
              Appendix (ACODA)
                Consonant (CNS)
```

Accent groups are the domains of pitch accents [6]. Features on the nodes of this structure express the phonological contexts that select which rules of interpretation are selected and fired. Nodes also form a dominance hierarchy and the 'headedness' of nodes affects their interpretation. There is no rule-ordering, nor any deletion of information present in the phonological structure. The interpreted structure is then converted to a parametric form depending on the signal generation method. The phonetic descriptions and timings can be used to select diphones and express their durations and pitch contours for output with the MBROLA system [1]. The phonetic details can also be used to augment copy-synthesis parameters for the HLsyn quasi-articulatory formant synthesizer [4]. The timings and pitch information have even been used to manipulate the prosody of natural speech using PSOLA [2].

3. USE OF EXTENSIBLE MARK-UP LANGUAGE

The Extensible Markup Language (XML) is an extremely simple dialect of SGML (Standard Generalised markup Language) the goal of which is to enable generic SGML to be served, received, and processed on the Web in the way that is now possible with HTML. XML is a standard proposed by the World Wide Web Consortium of industry specific mark-up for: vendor-neutral data exchange, media-independent publishing, collaborative authoring, the processing of documents by intelligent agents and other metadata applications [11].

We have chosen to use XML as the external data representation for our phonological structures in ProSynth. The features of XML which make it ideal for this application are: storage of hierarchical information expressed in nodes with attributes; a standard text-based format suitable for networking; a strict and formal syntax; facilities for the expression of linkage between parts of the structure; and readily-available software support.

In the ProSynth system, the input word sequence is converted to an XML representation which then passes through a number of stages representing phonetic interpretation. A declarative knowledge representation is used to encode knowledge of phonetic interpretation and to drive transformation of the XML data structures. Finally, special purpose code translates the XML structures into parameter tables for signal generation.

In ProSynth, XML is used to encode the following:

Word Sequences
The text input to the synthesis system needs to be marked-up in a number of ways. Importantly, it is assumed that the division into prosodic phrases and the assignment of accent types to those phrases has already been performed. This information is added to the text using a simple mark-up of: IP and AG.

Lexical Pronunciations
The lexicon maps word forms to syllable sequences. Each possible pronunciation of a word form has its own entry comprising: SYLSEQ (i.e. syllable sequence), SYL, ONSET, RHYME, NUC, ACODA, CODA, VOC and CNS nodes. Information present in the input mark-up, possibly derived from syntactic analysis, selects the appropriate pronunciation for each word form.

```
<AG HEAD="Y" START="0.5011" STOP="0.9727" STRENGTH="STRONG" WEIGHT="HEAVY">
 <FOOT HEAD="Y" START="0.5011" STOP="0.9727" STRENGTH="STRONG" WEIGHT="HEAVY">
  <SYL FPOS="1" RFPOS="1" RWPOS="1" START="0.5011" STOP="0.9727" STRENGTH="STRONG"
   WEIGHT="HEAVY" WPOS="1" WREF="WORD3">
   <ONSET START="0.5011" STOP="0.6516" STRENGTH="WEAK">
    <CNS AMBI="N" CNSCMP="N" CNSGRV="N" CNT="Y" NAS="N" RHO="N" SON="Y" START="0.5011"
     STOP="0.6516" STR="N" VOCGRV="N" VOCHEIGHT="CLOSE" VOCRND="N" VOI="Y">l</CNS>
   </ONSET>
   <RHYME CHECKED="N" START="0.6516" STOP="0.9727" STRENGTH="WEAK" VOI="N" WEIGHT="HEAVY">
    <NUC CHECKED="N" LONG="Y" START="0.6516" STOP="0.9727" STRENGTH="WEAK" VOI="N"
     WEIGHT="HEAVY">
     <VOC GRV="Y" HEIGHT="OPEN" RND="N" START="0.6516" STOP="0.8620">a</VOC>
     <VOC GRV="N" HEIGHT="CLOSE" RND="N" START="0.8620" STOP="0.9727">I</VOC>
    </NUC>
   </RHYME>
  </SYL>
 </FOOT>
</AG>
```

Figure 2 Extract of XML mark-up of "lie" taken from the phase "It's a lie."

Prosodic Structure

Each composed utterance comprising a single intonational phrase is stored in a hierarchy of: UTT, WORDSEQ, WORD, IP, AG, FOOT, SYL, ONSET, RHYME, NUC, CODA, ACODA, VOC and CNS nodes. Syllables are cross-linked to the word nodes using linking attributes. This allows for phonetic interpretation rules to be sensitive to the grammatical function of a word as well as to the position of the syllable in the word.

Database Annotation

We have recorded a medium sized database of simple phrases to explore variation in timing and intonation with segmental and prosodic constituency. This database of recordings has been manually annotated and a prosodic structure complete with timing information has been constructed for each phrase. Tools for searching this database help us in generating knowledge for interpretation.

An interesting characteristic of our prosodic structure is the use of ambisyllabic consonants. This allows one or more consonants to operate in the coda position of one syllable as well as in the onset position of the next syllable. Example are the medial consonants in "pity" or "tasty". The use of ambisyllabicity simplifies the rules for the temporal calculation of nucleus duration. However to achieve ambisyllabicity in XML it is necessary to duplicate and link nodes since XML rigidly enforces a strict hierarchy of components.

A small extract of a prosodic structure expressed in XML is shown in Figure 2.

4. KNOWLEDGE REPRESENTATION

In ProSynth knowledge for phonetic interpretation is expressed in a declarative form that operates on the prosodic structure. This means firstly that the knowledge is expressed as unordered rules, and secondly that it operates solely by manipulating the attributes on the XML encoded phonological structure.

To encode such knowledge a representational language called ProXML was developed in which it is easy to express the hierarchical contexts which drive processing and to make the appropriate changes to attributes. The ProXML language is read by an interpreter PRX written in C which takes XML on its input and produces XML on its output. ProXML is a very simple language modelled on both C and Cascading Style Sheets (see [12] for more information).

A ProXML script consists of functions which are named after each element type in the XML file (each node type) and which are triggered by the presence of a node of that type in the input. When a function is called to process a node, a context is supplied centered on that node so that reference to parent, child and sibling nodes is easy to express.

A simple example of a ProXML script to increase the duration of a nucleus according to the post vocalic context is shown in Figure 3. It is based on Klatt duration rule 9 [7]. In this example, the DUR attribute on NUC nodes is set as a function of the hierarchical context in which the NUC node is found and as a function of the phonological attributes found on adjacent nodes. Note that the rules modify the duration attribute (*= means scale existing value) rather than set it to a specific value. In this way, the declarative aspect of the rule is maintained.

5. FUTURE DIRECTIONS

In ProSynth, XML is used for a wide range of purposes: for input mark-up, for linguistic representation during synthesis, for storage of the lexicon and for annotation of a database. We also have a syntactic parse of each database phrase which we shall soon incorporate in the XML strcuture. A new XML processing language ProXML is used to formulate declarative knowledge for phonetic interpretation.

```
/* Klatt Rule 9: Postvocalic context of vowels */
NUC {
    node coda = ../RHYME/CODA;                          /* reference coda */
    if (coda==nil)
        :DUR = 1.2;                                     /* empty */
    else {
        node cns = coda/CNS;                            /* first consonant in coda */
        if ((cns:VOI=="Y")&&(cns:CNT=="Y")&&(cns:SON=="N"))
            :DUR *= 1.6;                                /* liquid */
        else if ((cns:VOI=="Y")&&(cns:CNT=="N")&&(cns:SON=="N"))
            :DUR *= 1.2;                                /* voiced stop */
        else if ((cns:VOI=="Y")&&(cns:NAS=="Y")&&(cns:SON=="Y")&&
          (../..:STRENGTH!="STRONG"))
            :DUR *= 0.85;                               /* unstressed nasal */
        else if ((cns:VOI=="N")&&(cns:CNT=="N")&&(cns:SON=="N"))
            :DUR *= 0.7;                                /* voiceless stop */
    }
}
```

Figure 3 Example ProXML script

We have investigated the relationship between our use of XML and the proposals put forward in the Speech Synthesis Markup Language SABLE [10]. We feel that SABLE operates quite differently to our system. Specifically, SABLE addresses the mark-up of text at the highest level of a text-to-speech system, while we use XML to represent the linguistic content of the utterance. SABLE seems to confound the difference between linguistic structure and phonetic realisation by using mark-up to control low-level acoustic-phonetic parameters (e.g. fundamental frequency) expressed along with the orthography. Such a design prevents a synthesis system choosing the most appropriate means of realising intonation for a given linguistic purpose. Thus it is more sensible to mark-up the linguistic function of each textual component rather than to indicate that it should simply be 'emphasised'. If it is known that some text is 'new' information, or is in contrast to previous information or is information in dispute, then the synthesis system can determine the appropriate type of emphasis. We hope to pursue such a linguistic mark-up of input text within the ProSynth project.

ACKNOWLEDGEMENTS

This work is supported by the U.K. Engineering and Physical Sciences Research Council. The author is grateful for the comments of his colleagues on the ProSynth project at UCL, York University and Cambridge University.

REFERENCES

[1] Dutoit, T., Pagel, V., Pierret, N., Bataille, F., Van der Vreken, O. (1996) "The MBROLA project: towards a set of high-quality speech synthesizers free of use for non-commercial purposes" Proc. ICSLP'96, Philadelphia, vol. 3, pp. 1393-1396

[2]Hamon, C., Moulines, E., Charpentier, F., (1989) "A diphone synthesis system based on time-domain prosodic manipulations of speech", Proc. Int. Conf. Acoustics, Speech and Signal Processing, p238.

[3] Hawkins, S., House, J., Huckvale, M., Local, J., Ogden, R., (1998), "ProSynth: an integrated prosodic approach to device-independent natural-sounding speech synthesis", Proc. Int. Conf. Spoken Language processing, Sydney.

[4] Heid, S., Hawkins, S., (1998) "PROCSY: A hybrid approach to high-quality formant synthesis using HLsyn", Third International Workshop on Speech Synthesis, Jenolan Caves, Australia, p219.

[5] House, J., Dankovicova, J., Huckvale, M., (1999), "Intonation modelling in ProSynth: an integrated prosodic approach to speech synthesis", Int. Congr. Phonetic Sciences.

[6] House, J., Hawkins, S., (1995) "An integrated phonological-phonetic model for text-to-speech synthesis", Int. Congress of Phonetic Sciences, Stockholm, p2:326.

[7] Klatt, D., (1979) "Synthesis by rule of segmental durations in English sentences", Frontiers of Speech Communication Research, ed B.Lindblom & S.Ohman, Academic Press.

[8] Local, J., Ogden, R., (1997), "A model of timing for nonsegmental phonological structure." In Jan P.H. van Santen, R W. Sproat, J. P. Olive & J. Hirschberg (eds.) *Progress in Speech Synthesis*. Springer, New York. 109-122.

[9] Ogden, R., Local, J., Carter, P. (1999) "Temporal interpretation in ProSynth, a prosodic speech synthesis system", Int. Congr. Phonetic Sciences.

[10] Sproat, R., Hunt, A., Ostendorf, M., Taylor, P., Black, A., Lenzo, K., Edgington, M., (1998), "SABLE: A standard for TTS markup", Third International Workshop on Speech Synthesis, Jenolan Caves, Australia, p27.

[11] http://www.w3.org/XML/

[12] http://www.phon.ucl.ac.uk/project/prosynth.htm