Query Language for Access to Speech Corpora

Andreas Mengel, Ulrich Heid

University of Stuttgart, Institut für Maschinelle Sprachverarbeitung
Azenbergstraße 12, 70174 Stuttgart, Germany

Summary: With more and more speech corpora at hand the unit selection technique is a promising approach in concatenative speech synthesis. What is missing are models of optimal parameters that sufficiently describe utterances to be produced and their corresponding counterparts in collections of speech data. Prior to this, existing corpora have to be annotated on possibly relevant linguistic and signal levels. This paper deals with standards developed in the MATE project for the uniform annotation of speech corpora to be represented in XML and a query language which can access these corpora. These standards may accelerate the identification of optimal elements for the annotation and description of parameters relevant for the unit selection technique.

INTRODUCTION

Recently, the unit selection approach to speech synthesis has gained more and more attention. So far, we have no full understanding of the relation between the shape of the speech signal itself and the factors influencing it. Neither all the factors have been identified nor has their influence on the spectral and temporal shape of the speech signal and techniques for its effective manipulation been fully discovered.

The unit selection approach is a technical solution to overcome this information gap: For a given part of an utterance to be synthesized, that unit out of a speech data base is chosen, that - according to the model - fits the actual situation best. Thus, although the unit selection approach is a technique born out of a lack of information in order to produce more naturally sounding synthetic speech, it could also be used for the identification of those linguistic and non-linguistic features that influence the characteristics of the speech signal. The following paper deals with the question what is needed for this and how it can be done.

One aim of the unit selection approach -similarly to other speech synthesis techniques - is naturally sounding synthetic speech. For the success of this technique will it become more and more important to set up and maintain large speech corpora and to provide technical solutions for rapid access to the segments to be selected. However, another crucial factor has rather been neglected, so far: it is a further aim and the actual rationale of the unit selection approach to select chunks of speech signals as large and fitting the actual utterance or signal situation as well as possible. But what are the properties that determine the appropriateness and thus make the unit chosen fit best? In order to identify these, more databases will have to be produced and labeled on various levels.

CATEGORIZATION OF SPEECH DATA

Although there are quite some data available, most speech databases are only labeled on phone, word or sentence level. Even worse is that the format of the annotation varies a lot (see for example (1) and (2)). Thus, for the exchange of the existing data and their use in different environments, new parsers have to be written to make the data accessible. In order to change this situation, the MATE project (EU Telematics Project LE4-8370) aims at a standard for the representation of the annotation of speech and language. MATE can be seen as following the TEI guidelines. MATE develops standards for the uniform encoding of descriptions of linguistic and speech data (3); MATE does not primarily aim at the standardization of linguistic data or theoretical approaches, but only at their encoding. The standard lists basic requirements for the annotation, proposals for the optimal representation of category information and provides sample annotation schemes for prosody, morphosyntax, coreference, communication problems, as well as methods to handle cross-level phenomena. The project aims at integrating different theoretical frameworks or other annotation approaches for these fields (4) within a common encoding format (5, 6, 7)

THE MATE APPROACH TO ANNOTATION

As MATE can be seen in close neighbourhood of the TEI guidelines (7), XML (6) will be used as a universally applicable formalism for the representation and encoding of the data.

Phenomenon independence: It does not matter if the units to describe are words, phones, prosodic features, perceptual or spectral items, gesture data or tongue positions. One homogenous data model can be applied for the uniform representation of any phenomenological description.

Theory independence: No restrictions are made with respect to theoretical conceptualizations of phenomena to be encoded. During the project, the markup is applied to the areas of prosody, morphosyntax, dialogue acts, coreference, and communication problems. All phenomenological descriptions can be encoded alike and be used in parallel.

Integrative approach: A set of recommendations will be formulated to ensure that information for different phenomena coded in different XML files can be accessed simultaneously and related to each other by using one parser. Wherever possible, description data are related to each other in the markup, e.g. by inheritance of time information.

Platform/application independence: Because the data are represented in XML and thus in ASCII, no special decoding software is needed. As XML becomes a standard in more and more environments (cf. (8, 9)), this will enhance the acceptance of the standard. Along with the standard, a tool is developed in MATE that reads data encoded in the standard, represents and displays them, allows the annotation of data, and writes the information in the format proposed. Additionally, filters for the conversion of prominent representation standards (e.g. the xwaves xlabel (1) notation or the BAS PARTITUR format (2)) are implemented. Furthermore, a query tool (Q4M) is integrated into the MATE software environment. All MATE software is implemented in Java to support linguistic and speech technology research on any platform.

QUERY LANGUAGE

Description	Example	Operators	Explanation
Comparison of elements by the values of their attributes			
to a string	($a.pos ~ "N")	~ !~	equals, does not equal
to a numerical value	($a.start < 0.2)	< <= > >= == !=	less, more, equal, not equal
relative to a other values			
as a string	($a.pos ~ $b.pos)	~ !~	equals, does not equal
as a numerical value	($a.end > $b.end)	< <= > >= == !=	less, more, equal, not equal
and a change	($a.f0 > $b.f0 * 2)	+ - * /	(mathematical operations)
position relative to other elements			
in a hierarchy	($a ^ $b)	^	is parent of
in a sequence	($a << $b)	, <<	is direct/any left neighbor of
related to time	($a [[$b)	% [[ ]] []][// @	(time relations)
membership of a set of			
elements	($a { $b)	{ !{ {}	is member of, no member of, union
attribute values	($a.pos { $b.pos)	{ !{ {}	is member of, no member of, union

Negation ("!") of single expressions and the combination of query expressions by logical operators ("&&" and "||") is also supported.

FIGURE 1. Overview of operators available in Q4M.

In the Q4M query language, the context of data may not only be determined by neighbouring units (e.g. other sounds), but by any other unit on any level. This general level independent conception allows for powerful retrieval of constellations of signal and linguistic phenomena (Figure 1). Definitions of queries in Q4M must be seen as constellation descriptions. The output of queries are tuples of those units that satisfy the constraints defined for the constellation. Result tuples are represented in XML. Elements of the result tuples are hyperlinked to the elements found in the queried documents. Complex query expressions and their results can more easily be analyzed and the results - XML-texts themselves - can be used for further and more constrained queries. At last, the output can be used as a new document describing phenomena relevant for speech synthesis. Figure 2 provides examples for two XML texts to be searched, a query expression and the result of a query represented in XML.

CONCLUSION

The application of multilevel annotation has two aims: First, the investigation of

Input files: snd.xml

```
...<phon id="ph41" type="f" start="1.231" end="1.289"/>
   <phon id="ph42" type="I" start="1.289" end="1.374"/>
   <phon id="ph43" type="t" start="1.374" end="1.431"/>...
```

tob.xml

```
...<tobi id="to23" type="H*"  at="1.321"/>
   <tobi id="to24" type="L*H" at="3.209"/>...
```

Query: *Sounds that contain an H* ToBI label.*

```
   ($p:phon)($t:tobi) ; ($p @ $t) && ($t.type ~ "H*")
```

Query result:

```
...<qutup id="q23">
     <el id="el45" href="snd.xml#id(ph42)">
     <el id="el46" href="tob.xml#id(to23)">
   </qutup>...
```

FIGURE 2. Input files, query, and query result in Q4M.

linguistically definable influences on the speech signal, namely those properties that make the speech signal sound appropriate in a given communication situation. The linguistic, acoustical, and perceptual entities taking part in this process have to be found and reduced to the minimal set of properties needed. Secondly, this set of information will also be the relevant annotation information for an effective retrieval of those units that fit best in a given synthesis situation.

On the way to the identification of these properties more refined and multilevel annotation has to be applied to speech data. In order to make these data uniformly accessible, a common standard and methods for retrieval of any linguistic situation in the data are crucial.

It is hoped, that by finding the relevant and necessary determinants of speech situations and signal properties also the relevant properties to be manipulated for the acceptable synthesis of speech will be found, thus making the unit selection approach dispensable.

ACKNOWLEDGMENTS

The work presented here was carried out in the framework of the MATE project (a project in the Telematics Application Programme of DG XIII E of the European Commission). The query processor and the MATE software environment are a joint effort of the members of the MATE project.

REFERENCES

(1) Entropic, xwaves/xlabel, http://www.entropic.com/products/esps/esps.html

(2) Schiel, F.; Burger, S., Geumann, A., Weilhammer, K., The Partitur Format at BAS, in *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain, II, 1295-1301, 1998.

(3) Dybkjaer, L., Bernsen, N.O., McKelvie, D. And Mengel A. The MATE Markup Framework. *MATE Deliverable D1.2*, 1998.

(4) Klein, M., Bernsen, N.O., Davies, S., Dybkjaer, L., Garrido, J., Kasch, H., Mengel, A., Pirrelli, V., Poesio, M., Quazza, S. and Soria, S., Supported Coding Schemes. *MATE Deliverable D1.1*, 1998.

(5) Goldfarb, C., *The SGML Handbook*, Oxford, PUBLISHER, 1990.

(6) World Wide Web Consortium, *Working draft: Extensible markup language (XML) version 1.0 part 1: Syntax*. http://www.w3.org/TR/REC-xml, 1998.

(7) Sperberg-McQueen, C.M., Burnard, L. (Eds.): *Guidelines for Electronic Text Encoding and Interchange*. TEI P3. Text Encoding Initiative. ACH, ACL, ALLC, Chicago, Oxford, 1994.

(8) Sproat, R., Taylor, P., Tanenblatt, M., Isard, A., A markup language for text-to-speech synthesis, in *Proceedings of the Fifth European Conference on Speech Communication and Technology*, Rhodes, 1997.

(9) Sun Microsystems, Inc., *Java Speech Markup Language specification*, http://java.sun.com/products/java-media/speech, 1997.