Preliminary Work on Speech Unit Selection Using Syntax-Phonology Interface

Sabrina Tiun, Tang Enya Kong
Unit Terjemahan Melalui Komputer (UTMK), Universiti Sains Malaysia, Penang, Malaysia
sab@cs.usm.my

Abstract

This paper proposes an approach which uses a syntax-phonology interface to select the most appropriate speech units for a target sentence. The selection of the speech units is done by constructing the syntax-phonology tree structure of the target sentence. The construction of the syntax-phonology tree is adapted from the example-based parsing of UTMK machine translation. In the process of constructing the syntax-phonology tree, we first identify the related trees from the speech corpus. Then, the generated subtrees based on the related trees are combined. The concatenation of the combined subtrees nodes is the synthetic utterance of the target sentence.

I. Introduction

The use of linguistic tree structure like syntactic tree, prosodic tree or phonological tree has been employed to improve the Text-to-Speech System (TTS). It is especially used in order to generate a more natural and fluent synthetic utterance. Before the birth of corpus-based or unit selection speech synthesis, the linguistic tree structure was used in prosody prediction. The common approach was the mapping of syntactic tree into performance trees [4][5][3]. In corpus-based speech synthesis, the selection of appropriate speech units was guided by using phonology tree as in [6] [10].

In this paper, we propose to select appropriate speech units for a target sentence using a representation tree structure of syntax and phonology merged together. As mentioned in [7], this kind of approach is based on an idea that speech unit retrieved from appropriate context is likely to be the appropriate speech units. In our study, the appropriate context will be the syntactic structure with matched phonetic form. By be able to parse a syntactic tree with matched word, speech units with correct prosodic information can be retrieved.

Based on the idea mentioned in the previous paragraph, we propose a model (Figure 1) that will take a phonemised sentence as an input and parses it into a dependency-based syntactic tree structure with prosodic information annotated at every node. The aligned speech units at the nodes of the constructed syntax-phonology tree will be the appropriate speech units for the target sentence. Therefore, synthetic utterance of the target sentence is generated by concatenating these aligned speech units.

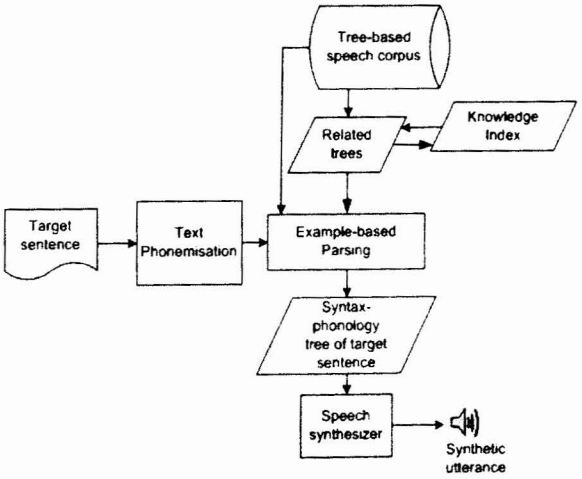

Figure 1. A proposed model of corpus-based speech synthesis using syntax-phonology interface

II. The Syntax-Phonology Tree Structure

The theory of transforming or mapping syntax structure into phonology [8] has evolved into collapsing syntax and phonology structure in one representation structure [9]. Whether it is about merging or integration of syntax and phonology together, it is all about using one abstract representation where both information on syntax and

978-1-4244-1692-9/08/$25.00 ©2008 IEEE

phonology existed. This is the type of representation tree we meant by a syntax-phonology.

In this paper, our tree representation is based on a dependency-based syntactic tree structure and prosodic information (which is a break index) integrated to the tree nodes. The nodes of this tree are phonetically transcribed so that we can match them against the phonemised words of the target sentence.

Our preliminary investigation on the synthetic speech production using this approach is by taking the Malay translated sentences from [2]. The sentences are recorded and parsed into dependency trees. Based on the recorded files, the tree nodes are segmented into words and annotated with break index (based on ToBI labeling schema). We meant by break index is the type of juncture between two words and in ToBI, the value is ranging from 0 to 4. The transcription and labeling of the break index are done manually by listening to the recorded sound of the translated sentences.

In order to ease the understanding of the syntax-phonology tree construction (or for the next sections, we will use the term tree combination), the dependency trees nodes in the next sections will be written orthographically and not in phonetic form.

For creating the speech database, the segmented word units are aligned with their corresponding tree. These aligned word units will be the speech unit candidates for the concatenation process.

In the subsection A, we will describe briefly about UTMK example-based parser. Afterwards, in subsection B, we explain how the parsing algorithm has changed due to the prosodic information.

A. The UTMK example-based parser

The UTMK Example-based parser is used in UTMK Example-based Machine Translation (EBMT) to parse an input sentence into a dependency-based representation tree structure. As claimed by [2], the representation structure of an input sentence is constructed by imitating the structure of similar sentence in the example-based corpus. However, since most of the time similar sentence will not be available, therefore another option is to choose sentences related to the input sentence. The related sentences will be obtained from the example-base corpus based on sentences that contain the same words. These related sentences will be used to construct a knowledge index. Then, the knowledge index will be used as information to generate subtrees. Following the three steps: (1) Distance calculation, (2) Normalization and (3) Replacement, in sequence, the generated subtrees are combined into a single representation tree structure. Detail on UTMK example-based parsing algorithm can be read further from [1] and [2].

B. The speech unit selection and tree combination

In [1][2], the related examples to an input sentence are based on examples that contain one or more same words. However, with the availability of the break index value, the conditions of retrieving related trees is not longer the same. The word order of target sentence and the corresponding nodes must be matched accordingly. For example, if the target word located at the end of the sentence, then its corresponding node must be with the break index value of 4. The break index 4 here indicates of a full intonation and therefore, corresponding word must be the last word.

Suppose we want to produce a synthetic utterance of a target sentence "***orang tua itu yang minum minuman keras kutip kotak biru itu*** " ("the old man who drinks alcohol picks the blue box up") [2]. Using a set of Malay sentences in [2] and also an added sentence of our own, figure 2 to figure 6 are the set of sentences and their corresponding syntax-phonology tree structures (together with their aligned speech units) representing our tree-based speech corpus.

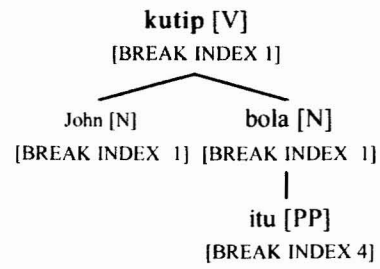

Figure 2. Syntax-phonology tree of sentence "*John kutip bola itu*"("John picks the ball up")[2]. (The word 'John' is corresponding to node 'Peter' at its corresponding tree structure in [2] and this must be a typo mistake since 'John' is the only object in this sentence.)

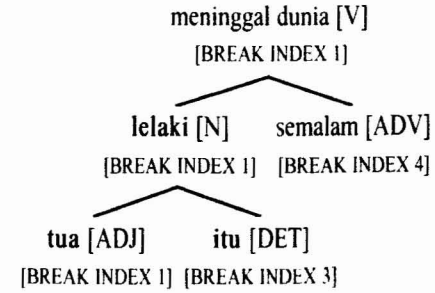

Figure 3. Syntax-phonology tree of sentence "*lelaki tua itu meninggal dunia semalam*" ("the old man died last night")[2]

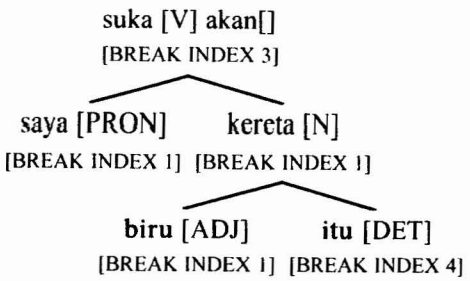

Figure 4. Syntax-phonology tree of sentence "*saya suka akan kereta biru itu*" ("i like the blue car")[2]

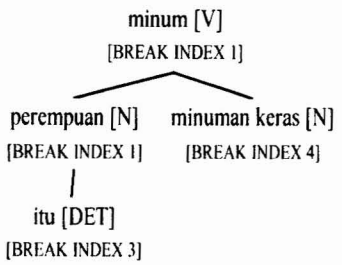

Figure 5. Syntax-phonology tree of sentence "*Perempuan itu minum minuman keras*" ("*that girl drinks alchohol*") [2]. ('*perempuan*' was spelt as '*permpuan*' in [2].)

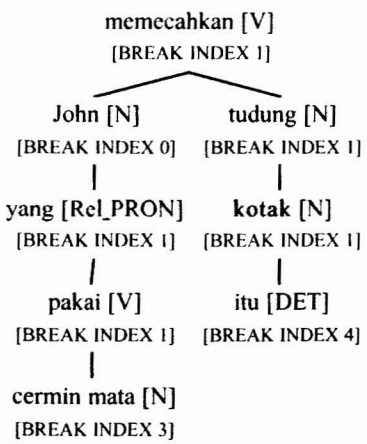

Figure 6. Syntax-phonology tree of sentence "John yang pakai kaca mata memecahkan tudung kotak itu" ("John who wears glasses breaks the cover of the box") [2]. (This tree and tree in [2] are not identical because we disagree with the word 'pakai' (' wear') as the head for a string "yang pakai cermin mata")

With the addition of break index to each node, related tree 4 will not be considered as structurally similiar with the target sentence. This is because the word '*minuman keras*' has a break index value of 4 but the word is not located at the end of the target sentence.

The break index value is given a highest priority in substituting or adjoining the trees. For example, if the tree contains a node of break index 4, it cannot longer permits a tree to be subtitute beneath it. And if break index 3 existed in one of the nodes in a tree, this tree cannot be substituted as the last order of subtree. Therefore, from the set of the related trees, the outcome of the trees combination is the syntax-phonology tree structure of the target sentence as shown in Figure 7.

Based on the syntax-phonology tree in Figure 7, the speech units chosen for concatenation are:

- From Figure 2, the chosen node is '*kutip*'.
- From Figure 3, the chosen nodes are '*lelaki*', '*tua*', '*itu*'.
- From Figure 4, the chosen nodes are '*biru*', '*itu*'.
- From Figure 5, the chosen node is '*kotak*'.
- From figure 6 the chosen nodes are '*yang*', '*minum*', '*minuman keras*'.

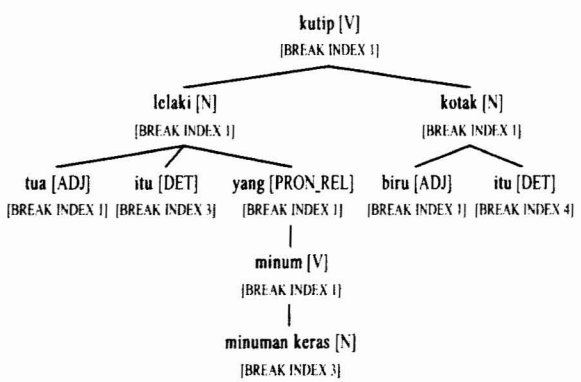

Figure 7. The syntax-phonology tree structure of the target sentence

III. PROOF-OF-CONCEPT BY EXAMPLE

Even though this work is using syntax structure but based on the prosodic information annotated at each of the tree node, a prosodic tree structure can be derived. Using break index types of 0, 3 and 4 as cues for prosodic words grouping, a prosodic tree structure of the target sentence can be constructed (Figure 8). The corresponding speech units of tree nodes in Figure 7 are concatenated without using any signal processing. However, silence unit is inserted at the right edge of each speech unit in which, its corresponding node tree has a break index type of 0 or 3. For the purpose of this investigation only, the concatenation is done

manually using Praat software (downloaded from http://www.praat.org)..

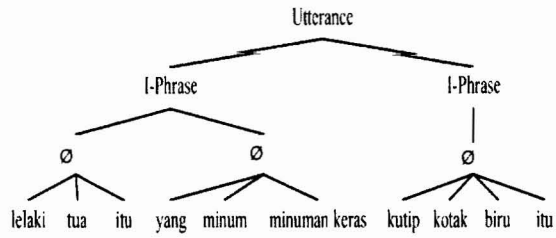

Figure 8. The derived prosodic structure from the syntax-phonology tree structure of the target sentence

We also record the target sentence and compare it with the synthetic utterance. By normalizing the time of both utterances, a comparison of smoothed pitch contours is drawn in Figure 9. Based on that Figure, both pitch contours look alike. When both utterances were perceptually tested, the locations of the phrasal break (break index type of 3) were at the same place.

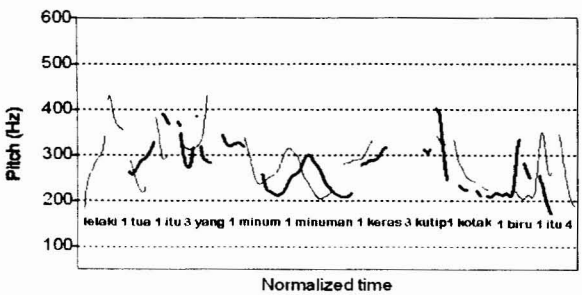

Figure 9. The comparison between a recorded utterance (*thin line*) and a synthetic utterance (*thick line*) pitch contours of the target sentence.

IV. Conclusion

This work is a preliminary investigation of selecting speech units using a syntax-phonology tree structure as an interface. Based on the preliminary result, we believe a high quality of Malay speech synthesis engine can be delivered and be implemented in a real environment of speech synthesis. Therefore, our next plan is to test this approach in a small limited speech synthesis system. Currently, the choice of speech unit is restricted at word level. However, we plan to improve this model further by adding smaller speech unit viz. syllable to the proposed model. This of course will need more prosodic information such as stressed and unstressed syllable annotated to the syntax-phonology interface.

References

[1] H.M. Al-Adhaileh and T. Enya Kong, "A Flexible Example-Based Parser Based on the SSTC". *Proceedings COLING-ACL*, 1998, pp. 687-693.

[2] H.M. Al-Adhaileh, *Synchronous Structured String-Tree Correspondence (S-SSTC) and Its Application for Machine Translation*. PhD Thesis, Universiti Sains Malaysia, Malaysia, 2002.

[3] M. Atterer and E. Klein, "Integrating Linguistic and Performance-based Constraints for Assigning Phrase Breaks". *Proceedings of COLING 2002*, 2002, pp. 29-35.

[4] J. Bachenko and E. Fitzpatrick, "A Computational Grammar of Discourse-Neutral Prosodic Phrasing in English". *Computational Linguistics*, MIT Press, Cambridge, 1990, pp. 155-170.

[5] L. Blin and M. Edgington, "Prosody Prediction Using A Tree-Structure Similarity Metric". *Proceedings of the Third International Workshop on Text, Speech and Dialogue (TSD 2000)*, 2000, pp. 369-374.

[6] A.P. Breen and P. Jackson, "Non-Uniform Unit Selection and the Similarity Metric within BT's Laurete TTS System". *Proceedings of the Third ESCA Workshop on Speech Synthesis*, 1998, pp. 373-376.

[7] B. Möbius, "Corpus-based Speech Synthesis: Methods and Challenges". *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung AIMS 6*, 2000, pp. 87-116.

[8] O.E. Selkirk, *Phonology and Syntax: The Relation between Sound and Structure*, *MIT Press*, Cambridge, 1984.

[9] M. Steedman, "Information Structure and the Syntax-Phonology Interface". *Linguistic Inquiry*. MIT Press, Cambridge, 2000, pp. 649-689.

[10] P. Taylor, "Speech Synthesis by Phonological Structure Matching". *Philosophical Transactions of the Royal Society*, Royal Society Publishing, UK, 2000, pp. 1403-1417.

Knowledge Profiling for Social Networking News Releases

Lo Tse-Yi[a], Vincent Khoo Kay Teong[b]

[a]*School of Computer Sciences,*
Universiti Sains Malaysia, 11800 Minden, Pulau Pinang, Malaysia
Tel : 604-6533888, Fax : 604-6573335
E-mail : tseyi.lo@live.com.my

[b]*School of Computer Sciences,*
Universiti Sains Malaysia, 11800 Minden, Pulau Pinang, Malaysia
Tel : 604-6533888, Fax : 604-6573335
E-mail : vkhoo@cs.usm.my

ABSTRACT

This paper presents a case study of the news item categorization for a social networking mashup. Since it is time consuming to develop the hierarchical news categories on an iterative and incremental basis using a conventional knowledge mapping approach, an alternative approach is derived by using the Subject Reference System (SRS) guidelines as the reference for categorizing various types of news releases. With the ulterior aim of enabling automatic generation of knowledge maps, this research project adopts only a manual process for the creation of knowledge profile entries, which would then form the input references for future automatic generation of knowledge maps. The approach for transforming a news item to a knowledge profile entry involves some computer-assisted extraction of keywords, which is then linked to the Subject Reference System (SRS) guidelines repository for ease of knowledge profiling. It is argued that the alternative knowledge profiling approach is easier and more efficient than a manufacturing-oriented conventional approach, because the knowledge ontology and validation of news releases have already been proven and accepted by a large community of news agencies.

Keywords

Knowledge Mapping, Knowledge Profiling, News Metadata, Knowledge Ontology, Automatic Knowledge Map Generation

1.0 INTRODUCTION

Information presented in a non-meaningful context for an intended audience is one of the contributing factors for information overload. It has been shown that Office-based and PDF documents are the dominant file formats on the Intranet (Littlefield, 2002). With the revolution of the Web 2.0 technology, it is foreseeable that documents circulated through the Intranet would become a major part of the file distributions throughout the Internet. Similarly, large amount of rich text documents available as news releases posted online would become one of the contributing factors for information overload. Information can only be transformed to knowledge when a proper context is specified. Thus, 'what constitutes knowledge' is constantly pondered upon by computer scientists and experts (Alavi and Leidner, 1999).

Information can be ambiguous without a proper context. A typical knowledge management challenge faced by organizations is not having a standardized approach for sharing and leveraging knowledge internally and externally (Liebowitz, 2004). Since the same set of information can convey different meanings in different context and presentation, one of the challenges is finding the most suitable domain for structuring them (Lê and Lamontagne, 2002). In this regard, knowledge mapping has been described as the process, methods and tools for analyzing knowledge domains in order to discover their inherent features and visualize them in a comprehensive and transparent form (Speel et al., 1999).