Lexicon and prosodic structure in a text-to-speech system

Sverre Stensby, Berit Horvei and Georg E. Ottesen
SINTEF DELAB, N 7034 Trondheim, Norway

ABSTRACT

This article is a presentation of a pronunciation lexicon and its use in improving prosody in a text-to-speech system. The inclusion of a lexicon is a natural way of giving synthesis systems information of word class and word pronunciation. The grammatical information is used to disambiguate homographs and to form a basis for prosodic structuring. The realized prosody is based on patterns for rhythmical units named feet and utilizes the information of word tone and pronunciation found in the lexicon.

INTRODUCTION

The scope of this paper is to present a pronunciation lexicon and its use in improving prosody in a text-to-speech (TTS) system. The paper addresses lexical entry, syntactic analysis, and prosodic structuring and realization. Parts of the work have been published previously, Stensby (1991, 1992, 1993).

LEXICON

When we read most of the words are known to us. We thus know their meaning and how to pronounce them. The inclusion of a lexicon in TTS-systems is a natural way of giving such systems similar information about the words in the text.

By lexicon we mean a dictionary in a machine readable form. Our lexicon is based on full word forms inflected by rule. Each entry includes the inflected word form, the pronunciation, the word class and grammatical function, and reference to the basic part of the word, i.e. the lexeme. The pronunciation is included because letter-to-sound rules will never be foolproof. For words not found in the lexicon, so-called unknown words, the pronunciation is found by rules.

Many complicated changes in pronunciation occur for the individual inflected forms of the same lexeme. These irregularities are the main reason for using a full word lexicon and not utilizing morphology in this first version. The method of inflection by rule allows for easy control of the produced words.

Basic word list

The basis for the lexicon is a small Norwegian dictionary and a list of the most common full word forms in newspaper text. The pronunciation and grammatical information given in the original dictionary are improved and extended. The grammatical information includes the word class and inflectional pattern for the word. The pronunciation includes indication of the stressed syllable and word tone. The inclusion of the word tone is important since there are no infallible rules for allocating this information.

We distinguish between open and closed word classes. The set of words belonging to the closed classes is mainly stable with time and is on the whole

covered by a finite number of words. Thus it is possible to include practically all words belonging to the closed classes in a dictionary. In this lexicon the set of closed classes are the infinitive marker, auxiliary verb, determiner, conjunction, preposition, interjection, and pronoun. These comprise in sum 525 words. The open word classes are noun, adjective, numeral, name, verb, and adverb. This gives jointly a basic word list of 12000 words, mainly consisting of lexemes.

Generation of inflected full word forms

Norwegian is partly an agglutinating language with many inflected forms created by adding suffixes. The inflected full word forms are generated by rule from the words in the base word list. Typically 4, 7 and 6 full word forms for nouns, adjectives and verbs respectively are generated, though there may be additional forms due to alternatives. To cover all these patterns 70 inflection models are included in the system. The resulting lexicon of inflected full word forms has 48000 entries. An example is shown in Table 1.

Table 1. *Example of generated full word forms for the verb grave (= dig). The apostrophes indicate the stressed syllable and the type of tone.*

Word	Pronunciation	Lexeme	Gram. category
grave	[gr'ɑ:və]	grave	infinitive
graver	[gr'ɑ:vər]	grave	present tense
gravde	[gr'ɑvdə]	grave	past tense
grov	[gr'u:v]	grave	past tense
gravd	[gr'ɑvd]	grave	perfect
gravende	[gr'ɑ:vənə]	grave	present participle
grav	[gr'ɑ:v]	grave	imperative

Table 1 illustrates that the change in pronunciation may be complicated even in words with regular inflection. Especially the word tone may change. The symbols ' and ` are used for tone 1 and tone 2 respectively.

Tones

Norwegian, Swedish, and some other languages have a limited use of tone, and most Norwegian dialects distinguish between two tones named tone 1 and tone 2. The difference is manifested in the stressed syllable where a fall in the fundamental frequency (F_0) is typical for tone 2. In Norwegian there are more than two thousand pairs of words with identical sound segments which are distinguished by tone alone. Many of these are homographs and need to be disambiguated.

The distribution of the tone 1 and 2 on the word initial syllable and a word internal syllable is shown in Table 2. Tone 2 predominantly occurs on the first syllable, while tone 1 is uniformly distributed between the initial and word-internal syllables. This information may be utilized in the task of allotting word tone to unknown words.

GRAMMATICAL ANALYSIS

The main task of the grammatical analysis is to disambiguate homographs and thus find the correct pronunciation and grammatical function. The system disambiguates

Table 2. *Relative occurrences of tone 1 and 2 in the lexicon depending on the accented syllable.*

Word tone	1	2
Word initial syllable	0.22	0.51
Word internal syllable	0.26	0.01

homographs of different word classes due to differences of their function in the sentence. The word class is used in assigning the prosodic structure.

In a text there may be words not found in the lexicon. Such unknown words are assumed to belong to one of the open word classes since all words belonging to the closed classes are presumed to be included in the lexicon. This restriction allows for treating the unknown words as homographs of the open word classes. The allocated word class may also be used by the pronunciation rules.

Parsing

The text is analysed in a multi-pass associative parser. In each pass the text is compared with phrase patterns, and a phrase is created when a match is found. The least ambiguous patterns are sought first, the prepositional phrase is an example of such a pattern.

PROSODIC STRUCTURING

The prosodic structuring utilizes the word class of the words, the established phrases, and rules for allotting sentence accent and focal accent. The prosodic structuring works locally within sentences. A unique division into sentences requires semantic analysis beyond the scope of this work. We therefore define every punctuation mark as a sentence divider, even though this gives a large number of sentences.

Focal accent (the highest prominence) is allotted to new information only, and is primarily located late in the sentence. This gives broad focus which we regard as the best when the assignment is based on simple assumptions only.

Focal accent is given to the last new noun or adjective in a sentence, if any, while ordinary accent is given to noun, adjective, verb or adverb. The word classes are given preference in the mentioned order. There is at most one accent per phrase and one focal accent per sentence.

The distinction between new and known is determined by a table of the lexemes occurring earlier in the passage. By use of the corresponding lexeme and not the word form itself, the words may be recognized in different grammatical forms. There is no forgetfulness in the present system, but this might be implemented simply by remembering the last few words of a passage, Horne (1992).

PROSODIC REALIZATION

The prosodic realization is based on patterns for F_0 and rhythm over accented and focally accented feet. A foot is a rhythmical unit starting with a pitch-accented syllable. The tone of the foot-initial word is assigned to the foot. The distinction between accent and focal accent is signalled by the F_0 at the end of the foot. The end of a foot coincides with the end of the text, or with the beginning of the next foot, or with a so-called foot-external section. The latter is a section which joins a

foot and a pause or two feet.

A comparison of F_0 and timing in an authentic sentence (solid line named Arne) and a synthetic sentence (dotted line named Modell2) is shown in Figure 1. The figure is generated by a system for studying prosody in texts read aloud, Ottesen (1992), Horvei et.al. (1993). In translation the sentence is "The painter worked for a month".

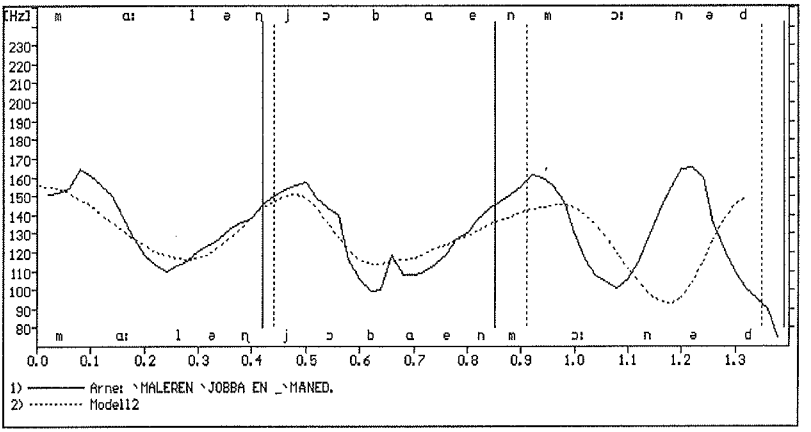

Figure 1. *Comparison of authetic and modelled fundamental frequency contour. Foot boundaries are indicated by vertical lines. The time scale is in seconds.*

The grammatical analysis, the prosodic structuring, and the prosodic realization are done by rule. The structure is indicated by vertical lines and apostrophs in Figure 1 and the _ indicates the focal accent. The established prosodic structure is identical to the reading, and there is good agreement between the authentic and synthesised contours apart from a slight difference in timing and in the final F_0.

ACKNOWLEDGEMENTS

This research has been funded by the Norwegian Telecom and was carried out in cooperation with the Department of Linguistics at the University of Trondheim.

REFERENCES

M. Horne (1992), "Semantic and pragmatic conditioning of accenting in Swedish - implications for speech synthesis", *Proc. Nordic Prosody VI, KTH, Stockholm.*

B. Horvei, G. Ottesen and S. Stensby (1993), "Analysing prosody by means of a double tree structure", *Proc. EUROSPEECH 93, Berlin, Germany.*

G. Ottesen (1992), "A method for studying prosody in texts read aloud", *Proc. ICSLP 92, Banff, Canada*, pp. 1271-1274.

S. Stensby (1991), "Prosody in a rule-based Norwegian text-to-speech system", *Proc. EUROSPEECH 91, Genova, Italia*, Vol 3, pp.1149-1152.

S. Stensby (1992), "Prosody in a text-to-speech system for Norwegian", *Proc. Nordic Prosody VI, KTH, Stockholm.*

S. Stensby (1993), "Struktur av prosodi i talesyntese", *Avhandling for graden doktor ingeniør, Institutt for teleteknikk, Akustikk. Norges Tekniske Høgskole, Universitet i Trondheim.* (In Norwegian. Abstract in English).