Automatic Vocabulary Extension for a Speaker-Adaptive Speech Recognition System based on CVC Units

Peter Fessler, Heidi Hackbarth, Marianne Kugler and Arnd Boehm

Standard Elektrik Lorenz AG, SEL Research Center Stuttgart, F.R. Germany

ABSTRACT

For speech recognition with large vocabularies, a user should not be burdened with having to train several thousand words explicitly. Therefore, it proves extremely useful to provide a means for easy vocabulary generation and enlargement from written text input. Applying a set of appropriately defined rules, the orthography of a lexicon item is first transcribed into the phonetic symbols of the standard pronunciation and the most common alternatives thereof. From these, the multiple sequence of specific subword units of a lexicon entry is produced. The tool introduced here is part of a comprehensive speech processing system for subword-unit based, speaker-adaptive recognition of continuous speech.

1. INTRODUCTION

A major goal of current speech recognition research is the identification of words from large vocabulary. Since whole-word recognition is not suited for vocabularies of several thousand words, the items in such a lexicon are encoded in specific subword units. This is bearing the advantage of a limited inventory of speech elements which only have to be pronounced by a speaker once (in various contexts), whereas vocabulary size is theoretically unrestricted.

A lexicon entry contains a list of several word descriptors, among which are the orthography and the subword-unit transcription of the standard as well as of the alternative pronunciation. Whereas the latter is usually inserted manually, for the large vocabularies we are investigating, it is preferable to derive the subword-unit transcription from written text automatically without any training by the user and with minor human interaction. This information is therefore independent of the pronunciation by an individual speaker. Thus, a lexicon can be referred to by any user, because variable pronunciations are coped with by alternative subword-unit transcriptions.

In this presentation, a tool is introduced for the automatic, principally unlimited vocabulary generation and enlargement from orthographic input. After the description of the comprehensive speaker-adaptive recognition system for continuous German speech of which the vocabulary extension tool is one part, the subword units used here are introduced in detail. Essential for successful operation is the compatibility between the subword-unit coding of the spoken utterance and the corresponding transcription of the orthographic input. Both analyses are explained in detail. Central to them is the lexicon with its entries in which also the automatically generated alternatives to the standard pronunciation due to intra- and inter-speaker variability of a spoken word are contained. This represents one attempt towards the incorporated speaker adaptivity, a broader description of which, however, is lying beyond the scope of this article. Finally, a summary will be given.

2. THE SPEECH RECOGNITION SYSTEM

Figure 1 is providing an overview over the recognition system for continuous speech from large, extendable vocabulary of several thousand words and more.

Intervals of 8 ms of digitized speech are passed through a 64 channel filter bank and transformed into 20 loudness functions, covering the frequency range from 50 Hz to 6.4 kHz. The speech signal is segmented into subword units. A detailed description of these subword units and the segmentation procedure will be given in the following sections. During a training phase, subword units are extracted from labeled speech data to build up the inventory. It is taken as reference when the segmented feature vector sequences of actual input are classified. The module dedicated to the word match thus has a lattice of multiple subword-unit hypotheses at its disposal, from which it generates overlapping word hypotheses. These, in turn, undergo a grammatical analysis that finally determines the spoken phrase.

Entries in the lexicon can be made by using the vocabulary extension module for written text which produces a transcribed subword unit string for the standard pronunciation of a word and alternatives which cover the most common variations in pronunciation. The result is a lattice of subword units that create a so-called word model. For successful recognition, it is necessary to gain consistency between this automatic transcription and the acoustical analysis of the speech signal. The description of the specific kind of subword units used here will help to explain this need for compatibility in segmentation.
3. SUBWORD UNITS

The word models in the lexicon are based on so-called CVC subword units. These three units per syllable are gained from feature vector sequences, segmented into vocalic syllable kernels surrounded by syllable-initial and -final consonant clusters (CVC). This type of subword units has been selected, because whole syllables, each spoken in various contexts, amount to a computationally untreatable inventory. Phoneme-sized units, on the other hand, may not be classified unambiguously without additional knowledge sources. In contrast, CVC units can be segmented and identified automatically and their inherent rigid CVC pattern per syllable is also advantageous at subsequent signal processing stages.

Due to coarticulation effects, one can further differentiate between syllable-initial consonant clusters in word-initial or word-medial position and syllable-final consonant clusters in word-medial or word-final position. The inventory thus comprises 5 partitions with 25 vowels, about 60 syllable-initial and some 210 syllable-final consonant clusters, whereby 150 elements of the latter cover several thousands of the most common German words.

Quite a few of the inventory items resulted from an update due to a deviating segmentation introduced by the vocabulary extension tool, in order to gain accordance with the processing of the acoustic signal.

5. CVC TRANSCRIPTION OF WRITTEN TEXT

The transcription of text from its orthographic form into CVC units is performed in three steps, the initial conversion of a graphemic string into a phonemic string with morph boundaries, second, the automatic generation of alternatives to the standard pronunciation, and finally the segmentation of the multiple phoneme string into CVC units, thus creating the word models of the lexicon.

5.1 Phonemisation

For the conversion of the graphemic string into a phonetic string containing morph boundaries, a modified version of SYNTESX was used (5). This software program is performing speech synthesis for German from written text. It consists of two parts, phoneme transcription and acoustic signal generation. Only the first part was used here. Its output is a string in broad phonetic
transcription approximately following the standard set by the German DUDEN pronunciation dictionary (6). Additionally, the morph boundaries relevant for the following syllable segmentation program are marked in the phonetic string. The encoding is done in CPA, a computer readable phonetic alphabet developed within the ESPRIT project 8607291 (7). The modifications of the conversion program mainly concentrate on an improved segmentation and phonetic rather than phonemic coding of vowels, i.e. two vowel qualities and two quantities are differentiated here, whereas for the original SYNTEX speech synthesis program usually only one quality exists. By thus extending the program the DUDEN pronunciation is produced. Some of the problems specific for German and relevant for the present task are discussed below.

Unlike in other languages, in German, words are frequently formed by derivation and concatenation, which can lead to very long words. Strong phonetic influences at these ‘word boundaries’ within words (equal to those across words) require an exact segmentation to provide a reliable basis for the conversion itself and the following syllable-oriented segmentation into CVC units.

A German word can be characterized recursively by the formula

\[
\text{WORD} = \text{(prefix stem (suffix) (ending) (Fuge) WORD
}\]

where (normally) up to two prefixes and suffixes can occur. ‘Fuge’ denotes a filling character in a composed word for ease of pronunciation. SYNTEx breaks down the word into letters or letter sequences using a cluster analysis. It is based on the fact that only a limited number of consonant sequences are allowed in morph initial (52) and morph final position (120). Any sequence not found must therefore be a complex sequence and can be split into a morph final and a morph initial basic sequence. Starting at the preliminary boundaries suggested by the cluster analysis, the algorithm searches for a complete parse through the word, applying rules for prefixes, suffixes and endings. For example, a complete subdivision of the word

\[
\text{‘BESTANDSAUFNAHME’ is ‘BE.STAND,S.AUF.NAHM,E’.
}\]

The full stop denotes a ‘hard’ morph boundary, i.e. a boundary which is phonetically characterized by devoicing of voiced plosives and fricatives, vowel reduction (‘e’ to schwa), use of the vocalic ‘r’ etc. ‘Soft’ morph boundaries (mainly between a stem and a following suffix or ending beginning with a vowel, here marked by the comma) do not invoke such coarticulation phenomena. With one exception, only the hard morph boundaries are relevant for the following segmentation. By making use of this information, finally a string of phonetic symbols (CPA), is encoded. For the segmentation rules according to the strong influence of certain phonemes on the position of the boundary of a CVC unit. For example, voiceless sounds and most of the voiced fricatives have a very low signal intensity, at least for low frequencies. Especially for /s/, /x/ and /z/, additional rules had to be established. Below, some examples are shown, in which the ‘-’ stands for a word boundary, ‘,’” marks the beginning of a word stem or a prefix, ‘,”’ indicates the beginning of a suffix or an ending. If the syllable boundary is put in the middle of a phoneme, the corresponding symbol is printed twice (e.g. the /z/ in "Prognose").

\[
\text{Landschaft lan-tSaft ("landscape")
}\]
\[
\text{Überschwang üb-schwan/s-vaN ("exuberance")
}\]
\[
\text{Hochzeit hoX-tsa/t ("wedding")
}\]
\[
\text{Binfall bi:n-faI ("idea")
}\]
\[
\text{Prognose prOg-noiz-z ("forecast")
}\]

Since the vowels are easily detectable the partitioning of the syllables into CVC units can now be performed.

Although the vocabulary extension module tool described above performs the transcription with a minor number of errors, corrections can be applied in an interactive mode. These refer mainly to foreign or loan words and names. Fortunately, it turned out that an encoding into CVC units already contained in the inventory is realizable in most instances.

5.3 Segmentation of Phoneme Strings

For the segmentation of the phonemic string of a word into CVC units, a set of rules has been developed that uses knowledge about the specific properties of the phonemes in different contexts. Initially, the rules for syllables of spoken German text proposed by Ortmann (8) have been applied, which we use for further CVC segmentation. Ortmann describes, e.g., the influence of morph boundaries on the position of the syllable boundary. Generally, a syllable is defined to start within the last consonant of a word-medial consonant cluster. Several exceptions to this default rule became necessary due to the alignment with the acoustic analysis. It became apparent that the Ortmann syllable structure in some cases conflicts with the articulatory and rhythmic pattern of speech without considering the contour of signal loudness as a main criterion, as required for our purpose. For this reason, we expanded the number of segmentation rules according to the strong influence of certain phonemes on the position of the boundary of a CVC unit. For example, voiceless sounds and most of the voiced fricatives have a very low signal intensity, at least for low frequencies. Especially for /s/, /x/ and /z/, additional rules had to be established. Below, some examples are shown, in which the ‘-’ stands for a word boundary, ‘,”’ marks the beginning of a word stem or a prefix, ‘,”’ indicates the beginning of a suffix or an ending. If the syllable boundary is put in the middle of a phoneme, the corresponding symbol is printed twice (e.g. the /z/ in "Prognose").

\[
\text{Landschaft lan-tSaft ("landscape")
}\]
\[
\text{Überschwang üb-schwan/s-vaN ("exuberance")
}\]
\[
\text{Hochzeit hoX-tsa/t ("wedding")
}\]
\[
\text{Binfall bi:n-faI ("idea")
}\]
\[
\text{Prognose prOg-noiz-z ("forecast")
}\]

Since the vowels are easily detectable the partitioning of the syllables into CVC units can now be performed.

Although the vocabulary extension module tool described above performs the transcription with a minor number of errors, corrections can be applied in an interactive mode. These refer mainly to foreign or loan words and names. Fortunately, it turned out that an encoding into CVC units already contained in the inventory is realizable in most instances.

5.2 Pronunciation Alternative

Due to inter- and intra-speaker variations in successive utterances, different pronunciations evolve. These have to be covered by the word models in the lexicon which represent the reference during the word match. Such alternatives are created automatically by a set of rules, applied to the CPA transcription of the standard pronunciation. The rules are derived from acoustic inspection of a large amount of speech data, recorded from many different speakers. The temporal synchrony between standard and alternative pronunciation is representing an intricate problem, in particular, when syllables are suppressed, because both have to fit the word models in the lexicon. This could be alleviated, however, by indicating those locations where a syllable is omitted.

6. LEXICON

A lexicon entry for a word consists of several items. Aside from orthography, the CVC
transcription of the standard as well as alternative pronunciations are noted. The number of syllables, the word class and stress marks are also included by interactive editing for use by successive modules. All entries are adjusted to a specific word model, in which the multiple CVC transcription of a word is represented in a very compact form with two linear variants per CVC unit and a further one for a syllable-size deletion (Figure 2a). This has the advantage of very fast lexical access and retrieval during classification as well as reduced size compared with each entry written in a full form.

<table>
<thead>
<tr>
<th>Orthography</th>
<th>Standard Transcription</th>
<th>Linear Variant</th>
<th>Deletion Variant</th>
<th>No. of syllables</th>
<th>Word class</th>
</tr>
</thead>
<tbody>
<tr>
<td>E R D</td>
<td>E R/t be:R Rn</td>
<td>-- Rt</td>
<td>-- -- Rn</td>
<td>3</td>
<td>noun</td>
</tr>
</tbody>
</table>

Figure 2 : Multiple transcription (CPA) of a lexicon entry with standard transcription, one linear variant and one syllable-size deletion (a). Realisation of this model for the German word "Erdbeeren" (= "strawberries") and part of its entire entry below (b).

7. SUMMARY

By alignment of the loudness segmentation and the text-to-CVC unit transcription, we obtained a very useful tool to create and extend large vocabularies.

With the help of this tool, we have created a lexicon of more than 8000 entries for our speech recognition system. This lexicon will be further extended in the future to cover declined and conjugated forms of the basic German vocabulary.

ACKNOWLEDGEMENT

The authors would like to thank Manfred Immendoefer for fruitful discussions on the theme and Michael Trompf for software encoding. The work described was partly supported by the German Ministry for Research and Technology (BMFT) under contract No. 413-5839-ITM880101. Only the authors are responsible for the contents of this publication.

References