Unit Selection Speech Synthesis for a Directory Enquiries Service

Stefan Breuer, Julia Abresch
Institute for Communication Research and Phonetics (IKP), Bonn University
E-mail: breuer@ikp.uni-bonn.de, abresch@ikp.uni-bonn.de

ABSTRACT

In cooperation with klickTel GmbH (Dorsten, Germany) a unit selection synthesis system for a directory enquiries application was developed on the basis of BOSS II (Bonn Open Synthesis System II, Stöber et al. 1999, Klabbers et al. 2001). The speech corpus used was designed and recorded specifically for the needs of proper name synthesis. In the following, we will give an overview of some of the steps taken for the adaptation of the synthesis system, as well as the corpus design and annotation.

1. INTRODUCTION

In the project described here, a speech synthesis system specialised on the output of names and addresses was developed at IKP. The application aimed at was an automatic directory enquiries service with the ability to present users with a synthesised acoustic output in response to their requests. The goal was to not only synthesise telephone numbers, but also the complete names and addresses of directory entries that match a given query input. Proper names pose a major problem for transcription and synthesis. From the viewpoint of corpus design they are problematic because the segmental transitions covered by standard reading texts for corpus recordings do not suffice. Proper names often do not follow the phonotactics of the standard variety of a language, either as a result of historical sound change or dialectal or foreign origin. Difficulties in transcription arise from the ambiguities and inconsistencies of orthographic conventions, or grapheme-to-phoneme mapping, within a language and even more so between languages.

2. BOSS II

BOSS II is a software architecture for non-uniform unit selection synthesis. In a unit selection system the building blocks for an output signal are selected at runtime from a large corpus of speech. In contrast to classical concatenative synthesis multiple units may (and should) exist in the corpus for each phone that is to be synthesised. Non-uniform means that the units used for a given sequence can be of different size and/or from linguistic levels. At the moment, BOSS II uses words, syllables and phones. Cost functions decide which of the alternatives for a given word, syllable, or phone is to be selected. These cost functions can be divided into unit costs and transition costs. Unit costs measure the deviation of a certain token from predicted values for duration, phrase position, lexical stress as well as differences between the canonical transcription and actual realisation. Transition cost functions on the other hand try to optimise the transitions between adjacent units, in the case of BOSS II by calculating spectral differences, using mel-cepstrum coefficients and F0. If a glitch occurs nonetheless, BOSS II employs F0 manipulation to smooth the transition. Likewise, duration is manipulated if it differs significantly from the predicted value.

3. TOKENISATION AND TRANSCRIPTION

Telephone directories contain a large number of abbreviations of different types, some of which have to be expanded to their unabbreviated form, while some need to be spelled (UN) or pronounced in their abbreviated form (e.g. NATO). To achieve a correct transformation, the following tokenisation strategy is employed: By means of a manually devised set of regular expressions, the most frequent abbreviations from our directory database are mapped to their unabbreviated counterparts. All words that still end with a period (.) or those that are written in capital letters after this first stage are checked for presence in an exception lexicon. If the lookup succeeds, they can be passed to the transcription as is. If not, they are checked for pronounceability against a regular expression describing orthographic syllables that can be pronounced in German. If there is a match there is another lookup in a list of pronounceable abbreviations that are usually spelled. Only if the abbreviation is found in this list, or if it does not match the regular expression, is it spelled.

Examples:

Dipl.	known abbreviation, expanded to “Diplom”
Chr.	ambig. abbrev. (Christoph/Christian?), spelled
DAGA	pronouncable abbrev., passed to transcription
SWB	unpronouncable, will be spelled
DAK	pronouncable, but not usually pronounced, will be spelled.

In BOSS II, transcriptions are provided by one of three distinct mechanisms, the first of which is a simple lexicon lookup, that uses BOMP (Bonn Machine-Readable

 ISBN 1-876346-48-5

Pronunciation Dictionary, Portele et al. 1995) as a source of transcriptions. If this fails, morpheme decomposition is attempted on the basis of a pronunciation lexicon of the most frequent German morphemes (Ortmann 1993). A cost function decides which of the possible decompositions is selected in case of an ambiguity. If the input word cannot be split into morphemes it is passed on to the last level of transcription, a decision-tree based grapheme-to-phoneme conversion. This last transcription module uses three different decision trees to sequentially map words to a transcription, syllabify the resulting string and then assign stress to it, similar to the method described in Busser (1998).

Over 45.000 first names, surnames, company names, street and place names from the directory database were transcribed in order of descending frequency during the course of the project. Combining these with BOMP results in a transcription database with about 245.000 entries. The name transcriptions alone cover a large portion of the name tokens contained in the directory database, although millions of types remain untranscribed.

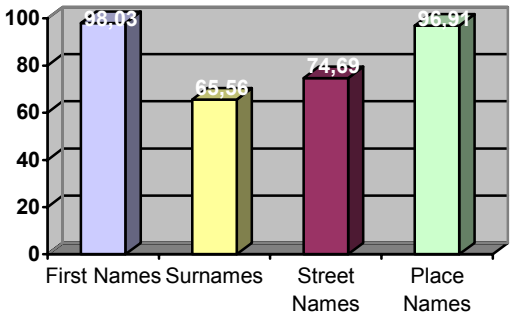

Fig. 1: Coverage of name tokens in the directory database by the 45.000 transcriptions of the most frequent names, in percent.

As there are significantly more surnames than e.g. place names, a much lower coverage of tokens is reached with approximately the same number of transcriptions.

The multilingual origin of the names in the directory database is problematic with regard to the definition of a proper phone set for transcriptions and the speech corpus. A level of nativisation[1], in this case, adaptation to the German phonology, has to be found that is a sound compromise between different, possibly conflicting conditions. Namely to be as close to the native pronunciation (in the language of origin) as possible, while still being a realistic representation of what an average German speaker could and would realise phonetically. If the transcription and the synthetic output are too authentic, it might be difficult for the user of the service to understand the name or to relate it to its orthographic representation. Additionally, the sound inventory needs to be limited to be able to keep the size of the corpus within certain bounds without cutting back on the coverage in terms of possible triphone contexts too much.

Based on the demographic proportions in Germany, the relevance of different languages for the name entries in the directory was estimated. For 15 languages (including English and French for company names), nativisation rules were defined to ensure a uniform transcription of grapheme strings by different transcribers.

The following sounds were added to the German phone set for the transcription of English names. The transcriptions are given in BLF (BOSS Label Format, see section 4.) which uses the SAMPA set of symbols:

/T/ as in *think*
/D/ as in *this*
/oU/ as in *nose*
/eI/ as in *bay*
/9:/ as in *service*
/O: / as in *dawn*

Other sounds were adapted to the German phonology, e.g. vowel/schwa diphthongs to vowel/vocalised-r [ɐ] [2] diphthongs

dare /d"e@/ /d"E:6/
fear /f"I@/ /f"i:6/

Some vowels were adapted in terms of quality and quantity, as in the following example:

Lady /"leI.dI/ /"leI.di:/

To ensure intelligibility for the average user of the system and to avoid an explosion of the number of possible phone-to-phone transitions, syllable-final devoicing of obstruents was extended to non-German words:

Edge /"EdZ/ /"?EtS/

For the same reasons, the glottal stop, that precedes syllable-initial vowels in German, was also used for the transcription of foreign words, as can be seen from the above example.

Some of the sounds that are part of the so-defined phone set are restricted for use with certain languages of origin. For

[1] Under nativisation we understand the process and the result of an adaptation of written or spoken foreign material to the phonology of the native language of a speaker in perception or production.

[2] This is the way coda-/r/ is usually realised in Standard German.

ISBN 1-876346-48-5

example, /T/ und /D/, are allowed for the transcription of English names, but it was decided not to use them for Spanish, because we assumed that German speakers are less familiar with their use in Spanish than in English

4. THE CORPUS

The experience gained from other projects led to the specification of a new set of synthesis units, named BLF (Breuer et al. 2001). In this set, liquids, glides and glottal sounds ([jhʔlʊwʀɹ]) are combined with following vowels and diphthongs to form new multi-phone units.

|#t|ra|n|s|.k|rI|p|.t|s|"jo:n|
(# = word boundary; . = syllable boundary; " = primary lexical stress)

These sounds exhibit a high level of coarticulation with the vowels they precede, which often renders them useless for concatenation in all contexts except for the one they were recorded in. The combination of these phones leads to a reduced number of unacceptable concatenations, but on the downside also to a significant increase of different base units.

The corpus designed for the project contains different types of texts to serve the various demands:

- a base set of sentences that covers a large portion of phone transitions for Standard German
- sentences containing about 2000 of the most frequent first names, surnames, street names, house numbers and place names in the appropriate syntactic positions, as specified by the output format of the synthesis (phrase slot filling)
- sentences containing numbers and pairs of numbers in different phrase positions
- spelled letter units, realised in different phrase positions

For a directory enquiries service, the adequate synthesis of telephone numbers evidently plays a central role. The prosody of telephone numbers is very complex and depends on a number of factors such as the number of digits, the repetition of single digits, or the degree of orderedness of a sequence (1234 vs. 3851), etc. (see Baumann et al. 2001). A relatively simple but effective measure to make the prosody of synthesised telephone numbers much more natural is the arrangement of the digits into phrases of maximally two units. If the number of digits is uneven, the first element is split off to form its own prosodic phrase. The rest are grouped into phrases of two digits, where the second element is realised with rising pitch unless it marks the end of the sequence of digits, in which case it is realised with falling pitch.

0228-45923 Null Zwei, Zwei Acht,
Vier, Fünf Neun, Zwei Drei.

To this end, the digits from 0 to 9 and all 100 possible pairs of 0 to 9 were recorded in different phrase-final positions. That way, each first element of a pair received medial pitch characteristics while the second elements were realised with either rising or falling pitch curves. In a similar way, dedicated units for spelling were created.

Additionally, special numbers that have a fixed prosodic realisation, such as the prefix for toll-free calls, "0800" were recorded separately.

0800: null-achthundert ('zero, eight-hundred')

This way, a significant improvement in terms of naturalness could be achieved in comparison to the often encountered systems that use a monotonous playback of single number utterances.

5. SUMMARY AND FUTURE WORK

Unit selection synthesis provides for a greater naturalness for the more frequent names of the database. On the other hand, it is hard to guarantee a certain degree of quality or intelligibility, when it comes to concatenating smaller units to synthesise the less frequent entries. This is why vital information, such as the telephone numbers or the output of the spelling mode has to be covered by dedicated units. An evaluation of the intelligibility of the system is currently in progress

To improve automatic transcription, we are planning to analyse the manually transcribed names to yield a list of morphemes or pseudo-morphemic components of names. This information can be used to augment the base list of German morphemes.

To judge the quality of our standards for the adaptation of foreign phones, it would be necessary to put the nativisation strategies applied by German speakers on an empirical basis. Experimental results for speakers of different L2 competence levels are needed. Unfortunately, little research has been undertaken in this field. Schaden (2001) investigated different pronunciation variants of place names to construct rules for the generation of alternative transcriptions for a speech recogniser. In contrast to that, for the application in speech synthesis, the aim would be to generate a single nativised variant that meets the expectations of the average native speaker of German.

ISBN 1-876346-48-5

ACKNOWLEDGMENTS

We would like to thank klickTel GmbH for funding the research on this project.

REFERENCES

[1] K Stöber., T. Portele, P. Wagner, W. Hess, "Synthesis by Word Concatenation" in *Proceedings Eurospeech*, pp. 619-622, Budapest, 1999.

[2] E. Klabbers, K. Stöber, R. Veldhuis, P. Wagner, S. Breuer, "*Speech synthesis development made easy: The Bonn Open Synthesis System*" in *Proceedings of Eurospeech*, Aalborg, 2001.

[3] T. Portele, J. Krämer, D. Stock "Symbolverarbeitung im Sprachsynthesesystem Hadifix" *Proceedings Elektronische Sprachsignalverabeitung*, pp. 97-104, Wolfenbüttel, 1995.

[4] W. D. Ortmann, *Kernmorpheme im Deutschen*. Goethe-Institut München, 1993.

[5] B. Busser, "TreeTalk-D: a machine learning approach to Dutch word pronunciation" in P. Sojka, V. Matousek, K. Pala, and I. Kopecek (Eds.) *Proceedings TSD Conference*, pp. 3-8, Masaryk University, Czech Republic, 1998.

[6] http://www.phon.ucl.ac.uk/home/sampa/home.htm

[7] Stefan Baumann, Jürgen Trouvain, "On the Prosody of German Telephone Numbers." *Proceedings of the 7th Conference on Speech Communication and Technology*, Aalborg, pp. 557-560, Denmark, 2001.

[8] Stefan Schaden, "Ein erweiterter Graphem-nach-Phonem-Umsetzer zur Modellierung nicht-muttersprachlicher Aussprachevarianten" in *Fortschritte der Akustik – Daga*, pp. 650-651, Bochum, 2002.