PHONETIC DATA BASES FOR GERMAN

K. J. Kohler

Institut für Phonetik und digitale Sprachverarbeitung
Kiel, Germany

ABSTRACT

The principles of automatically generating a large transcribed word corpus for German in a TTS environment, its extension to transcribed texts and the subsequent recording of an acoustic data base to be segmented and labelled are discussed.

1. SYMBOLIC, ACOUSTIC AND LABELLED ACOUSTIC SPEECH DATA BASES

In text-to-speech systems (e.g., RULSYS at KTH/Stockholm [1], a grapheme-to-phoneme module transforms orthographic text into phonemic transcription by rule. Since these rule-driven conversions are usually not optimal, a lexicon has to supplement the rule component, listing the exceptions that cannot be generated correctly. In order to be able to construct such a lexicon it is necessary to have a large representative corpus of words, linking orthographic forms with correct phonemic transcriptions. It is against such a corpus that the grapheme-to-phoneme output is evaluated: (1) the discrepancies between corpus and rule transcriptions have to be treated as exceptions and go into the lexicon, (2) the size of this lexicon, as a percentage of the total corpus, determines the efficiency of the rule system, (3) on the basis of the error types contained in the lexicon, an attempt can be made to improve the grapheme-to-phoneme rules and to reduce the exceptions lexicon, in a series of cycles, until an optimal balance is struck between lexicon size and number as well as complexity of the grapheme-to-phoneme rules, i.e. between the demands on storage and computing time, respectively.

Such a transcribed word corpus constitutes a symbolic speech data base, which - apart from being useful in TTS lexicon construction - can also be the basis for a great variety of phone statistics: frequency of phonemes, of phoneme sequences and clusters, of syllable types, of words containing a particular number of syllables, of simple and compound words (of various degrees of complexity), etc.. With the help of a powerful text-to-speech system it is possible to supplement the transcribed word corpus by a transcribed symbolic text data base (single sentences and whole texts), generated semi-automatically from orthography. The same type of statistics can then be carried out on running transcribed text, e.g., to establish the transitional probabilities in phonemic dyads, triads etc. through sentences.

Statistics at the symbolic phonetic level can guide the selection of word, sentence and text material for recording by a number of speakers, chosen according to a set of criteria (sex, age, social background, dialect etc.), to set up a representative acoustic speech data base for various purposes (basic research, improvement of the phonetic rules in a TTS system, training speech recognisers). To be optimally useful the acoustic data base needs processing: (a) segmentation, (b) alignment of transcription symbols with the demarcated signal sections, (c) various signal analyses (e.g. FFT) whenever necessary (labelled acoustic speech data base). Then a new type of statistics becomes possible which combines information about the symbolic and signal aspects of speech. We may, for instance, want to collect all the instances of the signal portions corresponding to a particular phoneme in the labelled acoustic speech data base.

2. SYMBOLIC SPEECH DATA BASE FOR GERMAN

2.1. Generating a Corpus of Transcribed Words

At the Kiel Phonetics Institute, a corpus of 23986 orthographic words was compiled from two frequency-ordered word lists. One contained the just over 9000 most frequent words from a computer survey of the newspaper 'Die Welt' carried out at the German Department of Lund University. As a newspaper style lacks many common words, such as 1st and 2nd pers verb forms, and therefore does not provide, e.g., 'bin, bist' (*(I) am, (you) are*), it was necessary to supplement this corpus to get a more comprehensive and representative coverage. So a second list was compiled from all the words in a wide spread of literary texts (fiction at different levels, various forms of journalism, legal administration texts, user instructions, but not scientific texts), available in ASCII format in the German Department of Kiel University, amounting to appr. 24000 frequency-ordered items. They were edited: spelling errors were corrected and most personal and topographic names as well as foreign loans excluded, except for very common ones, resulting in just over 21000 words. With the help of RULSYS support programmes, each word list was arranged in lines of five words and the lines consecutively numbered in steps of 5, starting with 0. Then the two lists, which had 6250 words in common, were combined to the total corpus of German words in freqency order. (Rolf Carlson at KTH carried out this amalgamation.)

This combined corpus was automatically transcribed by the grapheme-to-phoneme conversion module within the German TTS in the RULSYS environment, marking lexical stresses (' before the stressed vowel) and word boundaries in compounds (#), as well as affixing a general word class marker (w). The output was then manually corrected with regard to errors in phonemes, stresses, and word boundaries; at the same time function words were marked by + after the segmental string, and the general word class indicator changed to various subcategorisations in function words. This function word marking and classification is important for a syntactic analysis component within TTS. The corpus of German words thus has two related files, an (orthographic) text file and a (phonemic) transcription file, with identical item arrangements. The following examples illustrate the principle of corpus organization.

```
CORP.TX
  0 DER,DIE,UND,IN,VON.
  ...
110 SONDERN,IHRER,BUNDESREPUBLIK,
    NEU,HIER.

CORP.FO
  0 D'E:R+de,D'I:+de,'UND+bc,
    'IN+pp, F'ON+pp.
  ...
110 Z'ONDERNw,'I:RER+ns,
    B'UNDEZ#REPUBL'IKw,N'EUw,
    H'I:Rw.
```

As can be seen the transcription is very close to conventional spelling and quite abstract, i.e. morphophonemic rather than phonemic or phonetic, in that it keeps final voiced obstruents in order to preserve the relationships within inflectional paradigms, e.g, 'Haus' and 'Hauses' are both transcribed with Z at this level of abstraction although the former is ['haus], the latter ['hauzəs] at the phonetic realisation. Similarly, [ə] and [ɐ] are represented as E and ER, because the former can be derived from the latter by a lower-level phonetic rule. By applying these phonetic rules other, more-phonetic corpora can be derived, as, e.g., in

```
CORP.FON
  0 D'E:r+de,D'I:+de,'UNT+bc,
     'IN+pp,F'ON+pp.
  ...
110 Z'ONDrNw,'I:Rr+ns,
     B'UNDEOS#REPUBL"IKw,N'EUw,
     H'I:rw.
```

(Note the use of " for secondary stress and r for [ɐ].) By a simple conversion programme the phonetic notation can also be transformed into SAM-PA [5]:

```
CORP.SAM
  0 d'e6,d'i,'Unt,'In,f'On.
  ...
110 z'Ond6n,'ir6,
     b'Und@srEpUbl"Ik,n'OY,h'i6.
```

2.2. Generating an exceptions lexicon

The machine output from the TTS grapheme-to-phoneme module (CORP. MA) and its manual correction (CORP.FO) differ in those items that are not generated correctly by rule. They have to be included in the exceptions lexicon of the TTS system. This lexicon is generated automatically with the help of another RULSYS support programme that compares CORP.MA and CORP.FO (in relation to CORP. TX) and lists all the non-congruous items together with their orthographic versions in a new file, which is brought into alphabetical order by SORT. The following is an excerpt, providing, in each line, the orthographic word, the correct transcription, the machine transcription and a number referring to the frequency rank of the item in the corpus.

```
LEX.GE
AM 'AM+pp  * 'A:Mw ##  45
AMEISEN  'A:MEIZENw   *  AM'EIZENw
   ## 6452
AMERIKAREISE AM'E:RIKA:#R'EIZEw
   * AMERIKAR'EIZEw ## 6454
AMERIKAS AM'E:RIKA:Zw
   * 'A:MERIKAZw ## 2655
```

At present, this comparison yields 4150 errors, i.e. a rule efficiency of 82.7%. Changes in the grapheme-to-phoneme module can be quickly tested in their power of reducing the lexicon size by running the automatic programmes of corpus transcription, comparison and lexicon generation with reference to the transcribed data base CORP.FO, which is thus of fundamental importance for TTS development. But the procedure described so far has a serious flaw in that it includes all the items of an inflection or derivation paradigm in the lexicon, when there is a discrepancy between CORP.FO and CORP.MA, although the listing of a root would be sufficient to cover the exceptions of the whole paradigmatic set. This optimisation can be achieved by generating a corpus with suffix markings and root forms, applying a TTS suffix stripping module to CORP.TX and putting the result in the same format, as in

```
CORP.IN
150 MÜß,MÜß-EN,IHR-EN,FRAG,FRAG-E.
```

The generation of this inflected corpus is semi-automatic through further RULSYS support programmes and with manual correction of wrong suffix markings. Expanded CORP.TXR and CORP.FOR, containing the added roots, are also generated semi-automatically. The result is a corpus of 29183 items, i.e. 5197 roots have been added. The creation of the exceptions lexicon is then again fully automatic and yields 3754 entries at present,

i.e. a reduction of 9.5% compared with the original base not containing roots. This also means that the rule efficiency has been increased and that the TTS system has been made a great deal more general, allowing the correct generation of many more exceptions than are actually contained in the original data base. All that is necessary is the application of the same suffix stripping module in TTS processing of orthographic text input and subsequent lexicon look-up procedures, followed by a root modification module making phonemic adjustments in morphemic composition.

2.3. Generating corpora of transcribed texts

With the help of the German TTS system, developed and improved on the basis of an extensive word corpus, it is now possible to enlarge the symbolic speech data base for German and incorporate phonetically transcribed texts in addition to isolated words. Because the combined rule and lexicon efficiency in grapheme-to-phoneme conversion is very high, manual correction is minimal, and new transcribed texts can thus be generated from orthography more or less automatically. I have done this, starting with the standard sentences for German speech tests (Berlin (Sotschek) and Marburg, [4]) and two standard texts (The Northwind and the Sun; The Butter Story), illustrated in the following excerpt in adapted SAM-PA transformation (Q = [ʔ]):

```
001 h'OYt@ QIst S'2n@s
     fr'ylINsv"Et6.
002 di z'On@ l'Axt.
003 QAm bl'AU@n h'Im@l ts'i@n di
     v'Olk@n.
```

These transcribed data were then searched by appropriate programmes for all the phoneme dyads in German (including the transition to the first and from the last phoneme of a sentence). Frequencies of occurrence were entered into matrices, and empty cases that resulted from phonotactic restrictions and low phonotactic probabilities disregarded. For the remaining empty cases further sentences were constructed to cover all the 1308 most likely phoneme dyads in German by at least one instance. This resulted in a corpus of 398 sentences plus the two texts.

3. ACOUSTIC DATA BASE FOR GERMAN

The text materials in 2.3. were DAT recorded by 25 male and 25 female speakers, one in each group reading the whole corpus, the others various subsections of appr. 80 sentence equivalents, on average. The total recorded corpus comprises 4836 sentence equivalents, available in computer-readable form (16 kHz, 16 bit) on cassettes, with headers providing information about speaker, sentences etc. It will be segmented and labelled following the principles in [2,3]. The aim is to progressively automatise the processing.

4. REFERENCES

[1] CARLSON, R., GRANSTRÖM, B. & HUNNICUTT, S. (1990), "Multi-language text-to-speech development and applications", in *Advances in Speech, Hearing, and Language Processing*", Vol. 1, (W.A. AINSWORTH, ed.), London: JAI Press, 269-296.

[2] CARLSON, R. & GRANSTRÖM, B. (1985), "Rule controlled data base search.", *STL-QPSR*, *4*, 29-42.

[3] id. (1986), "A search for durational rules in a real-speech data base", *Phonetica*,*43*, 140-154.

[4] SOTSCHEK, J. (1976), "Methoden zur Messung der Sprachgüte I: Verfahren zur Bestimmung der Satz- und der Wortverständlichkeit", *Der Fernmelde-Ingenieur*, *30(10)*, 1-31.

[5] WELLS, J. C. (1987), "Computer-coded phonetic transcription", *JIPA*, *17*, 94-114.