AUTOMATIC GENERATION OF GERMAN PRONUNCIATION VARIANTS

Maria-Barbara Wesenick

Institut für Phonetik und Sprachliche Kommunikation (IPSK),
Ludwig-Maximilians-Universität München, Schellingstraße 3/II,
80799 München, Germany
wesenick@phonetik.uni-muenchen.de

ABSTRACT

The subject of this paper is a rule corpus of approx.1500 phonetic rules that models segmental variation of pronunciation in German connected speech. The phonetic rules express on a broad-phonetic level phenomena of phonetic reduction in German that occur within words and across word boundaries

The rule corpus has been designed as a component of the Munich AUtomatic Segmentation System (MAUS), which is an HMM-based system that produces the transcription of a speech signal and corresponding segment boundaries given the orthographic representation of the concerning utterance (refer to Kipp et al. [2] for details). The fact that speech is highly variable has been taken into account using the rules to complement the statistical modelling of German speech sounds and constrain the Viterbi-search.

In this paper first a short introduction to the phenomenon of variability of speech and our approach of dealing with this problem in a technical application is presented. This is followed by a formal description of the syntax of the rules and the inventory of symbols that is used. Finally, I give an outline of reduction phenomena in German and how they are represented in the phonetic rules.

1. THE REPRESENTATION OF SEGMENTAL VARIATION IN GERMAN IN PHONETIC RULES

A fundamental property of speech is that it is highly variable. No two utterances of the same word are ever produced exactly the same. Variability concerns the production of the same utterance of different speakers as well as the repeated production of an utterance by a single speaker (inter- vs. intra-speaker variability). Variability of speech depends among other factors on the immediate communicative situation, on the speechrate, speaking style and complexity of the semantic contents of utterances. Variability becomes apparent in the different realizations of a planned utterance. These can lie in a range from very clear, slow and precise to strongly reduced and fast. This scale is known as the hyper-hypo continuum of speech (Lindblom [4]).

Endeavouring to be intelligible and easy to understand for a listener speakers attempt on the one hand to speak clearly and with precise articulation. This conflicts on the other hand with the general tendency to keep the articulatory effort as low as possible. In order to compromise speakers permanently adjust their performance by taking into account the amount of information the listener can obtain from the communicative situation and the context of an utterance (system-oriented factors). Depending on the amount of information of the system-oriented factors the information that is contained in the speech signal itself (output-oriented factors) need to be more or less explicit. Hence, the expected variation of the utterances along the continuum of hyper- and hypospeech.

For many speech processing tasks a representation of the pronunciation of the language concerned is required which is usually taken from common pronunciation dictionaries. The main problem with this is that dictionaries mostly give only one possible form of pronunciation which is usually not the most common form. In speech technology and especially in the field of speech recognition the variability of utterances is difficult to deal with. A way to handle it is to grasp it in statistical word models. But if it is necessary to refer to smaller units on a phonemic or broad-phonetic level one has to take into account knowledge about phonetic processes that lead to the variability, because free phoneme-recognition has not been satisfactory yet. Concrete, complex and consistent information about possible variation in pronunciation is required. In segment-based speech recognition applications it is indispensable to process as much information about variation in pronunciation as possible for the analysis of the multifold input of human speech and the development of reliable systems in the area of speech technology.

The rule system is an attempt to grasp the different pronunciation forms of an utterance within the hyper-hypo continuum on a symbolic level taking into account articulatory processes. The citation form, that reflects the phonemic structure of a carefully pronounced single word, serves as a reference form from which all other hypothesized pronunciation forms can by derived by symbolic rules. Thus the rules describe in an abstract way segmental differences between the reference form and the pronunciation form that results form reduction.

The rules are based on knowledge about those articulatory reductions that have been observed in manual transcriptions and those that are reported in the relevant literature (e.g. [3]).

Our aim is to be able to generate with the rules all variants that may occur in actual speech of standard German. For this an empirical investigation of large databases of labelled speech material is necessary, because we do not yet know about all existing pronunciation forms that we want to model. But despite the fact that only an automatic procedure for labelling speech enables us to undertake empirical studies of segmental reduction to a sufficient extent (manual labeling is too expensive and time-consuming) we needed to find a way to provide enough rules for possible variants. Therefore, the corpus was designed so that in addition to naturally occurring variants also such pronunciation forms can be generated that might never occur.[1] This is necessary because a variant that we cannot describe can never be found with the system. On the other hand, wrong forms cannot compete with the correct forms during the Viterbi-alignment stage of MAUS. By using the automatic segmentation system and evaluating its output the rule corpus is successively being refined. So that gradually we obtain a set of rules that ideally reflects all and only naturally occurring reduction processes of German.

In its present form the rule corpus enables us to generate 94% of the pronunciation variants that result from manual transcription of the Phondat-II corpus of German read speech.

2. EXPRESSING SEGMENTAL VARIATION OF PPRONUNCIATION IN RULES

The actual pronunciation forms that result from articulation differences vary in every respect along a continuum because of the continuous nature of the reduction of articulatory movements. For a written description of the pronunciation forms the discrete symbols of a phonetic alphabet are used; the labelling level of MAUS and the description level of the rules respectively is a broad-phonetic level.[2] The symbols however, represent categories and reflect the existing continuum only relatively and in an abstract way.

Variation that is not explicitly contained in the rules or the generated variants is captured by the statistical HMM-models. The level of abstraction of the rules is a compromise between providing a maximum amount of phonetic information and using a very limited set of symbols for minimal complexity, that is aimed at in a statistical modelling of speech.

3. INVENTORY OF USED PHONETIC SYMBOLS

The phonetic symbols that are used in the rules are taken from the SAM-phonetic alphabet for German (as reprinted in Pompino-Marschall (ed.) [5]). The symbols stand for the phonemes of German.

Complete list of phonetic symbols used in the rules (respectively the corresponding HMMS):

- unreduced vowels: `a, a:, e:, I, i:, O, o:, U, u:,E, E:, 9, 2:, Y, y:`
- reduced vowels: `@, 6`
- diphthongs: `aI, aU, OY`
- plosives: `p, b, t, d, k, g, Q`
- fricatives: `f, v, s, z, S, Z, C, j, x, h`
- nasals: `m, n, N`
- liquids: `l, r`

Burst and aspiration of plosives are not marked. The phonological voiced-voiceless distinction refers to the difference in energy, namely the distinction fortis-lenis. Apart from the phonetic symbols the following special symbols are used:

- vowels: `!v`
- consonants: `!K`
- nasals: `!N`
- word boundary: `#`
- arbitrary word boundary: `&`

4. SYNTAX OF THE RULES

A rule r_i, i = 0...N-1 of the rule corpus consists of a right and a left part which are separated by “>”. The left part consist of a string of symbols $\mathbf{a_i} = < a_i\ (0), \ldots\ a_i\ (K_i - 1) >$, which has to correspond to a part of the canonic form of a word. The right part consists of a string of symbols $\mathbf{b_i} = < a_i\ (0), \ldots\ a_i\ (K_i - 1) >$, which stands for the variation of the string of symbols of the canonic form. a_i (k) and b_i (l), k = 0...K_i, l = 0...L_i are the symbols of the SAM-phonetic alphabet as listed in section 3.

Is one of the special symbols `!v`, `!K`, `!N` or `#` used on the left part of a rule, it also has to appear on the right part at the corresponding place. For `#` of the left part the symbol `&` can be used on the right side, if appropriate. Each rule is proceeded by a digit. It is used to group the rules.

Examples:

- `1nf>mf` — `nf` can be replaced by `mf`
- `1#pf>#f` — `pf` at the beginning of a word can be replaced by `f`
- `1g@n#>gN#` — `g@n` at the end of a word can be replaced by `gN`

1. That also means that we do not try to set up a complete and reliable phonological rule corpus, but a set of rules that is designed for practical purposes and as a phonetic component in a technical application for speech analysis.
2. In our terminology we follow Barry and Fourcin [1].

- 1t#t>&**t** — If two t meet at a word boundary they can be replaced by a single t. The word boundary then is arbitrary and is placed before the concerning segment.
- 1!vtp>!vQp — tp is replace by Qp after a vowel

5. DESCRIPTION OF TYPICAL PHONETIC PROCESSES IN GERMAN

In this section typical reduction phenomena of connected speech in German are described. These phenomena are all continuous in nature but described with discrete symbols on a broad-phonetic level. Examples for rules for the concerning reduction phenomenon are given and for illustration German words and their pronunciation that are liable to this type of reduction.

5.1. Assimilation

The assimilation of a segment to an adjacent segment in one or more parameters is a frequently observed phenomenon in connected speech. It affects the place or manner of articulation or the voicing parameter of a preceding segment (regressive) or a following segment (progressive). Assimilations can be total or partial and occur as well within morphemes or syllables as across them.

Assimilation of Place of Articulation

1. 1tp>p — <Mutprobe> (test of courage)
   mu:**tp**ro:b@ → mu:**p**ro:b@
2. 1tk>k — <mitkriegen> (get, catch)
   mI**tk**ri:g@n → mI**k**ri:gN
3. 1t#k>&k — <hat kurz> (has shortly)
   ha**t#k**U6ts → ha**&k**U6ts
4. 1b@n>bm — <geben> (give)
   ge:**b@n** → ge:**bm**
5. 1k@n>k — <trocken> (dry)
   trO**k@n** → trO**kN**

Examples 1. - 3. show a regressive assimilation of the place of articulation of the first plosive to the following plosive as the result of a continuous lenition process. In connection with an elision of the reduction vowel /@/ predominantly a progressive assimilation of place happens. This is shown in examples 4. - 5. where the place of articulation of the nasal involved is assimilated to the preceding plosive in connection with a @-elision.

Assimilation of Manner of Articulation

6. 1b@n>m — <Abend> (evening)
   Qa:**b@n**t → Qa:**m**t
7. 1gn>Nn — <Magnet> (magnet)
   ma**gn**e:t → ma**Nn**e:t
8. 1nd>n — <Verbindung> (connection)
   f6bI**nd**UN → f6bI**n**UN
9. 1st#d>&s — <hast du> (do you have)
   has**t#d**u: → ha**&s**u:

In 6. and 7. a regressive assimilation of manner is shown. This type of assimilation most frequently occurs when a nasal is involved. In 6. first occurs a @-elision (b@n → bn) and an assimilation of place (bn → bm) before the labial plosive is assimilated to the following nasal m (bm → m). In 7. the velar plosive changes to a velar nasal before the alveolar nasal.

Examples 8. and 9. show a progressive assimilation of manner where a segment is influence by a preceding segment in such a way that it assimilates to the manner of articulation of this segment. In 8. the alveolar plosive changes to an alveolar nasal, in 9. the alveolar plosive to an alveolar fricative.

Assimilation of the Voicing Parameter

10. 1pz>ps — <Absicht> (intention)
    Qa**pz**ICt → Qa**ps**ICt
11. 1t#d>&d — <hat der> (has the)
    ha**t#d**e6 → ha**&d**6
12. 1t#d>&t — <hat der>
    ha**t#d**e6 → ha**&t**6

This type of assimilation refers to the case that a usually voiced segment is pronounced unvoiced influenced by an adjacent segment that is itself voiceless. The reversed case is also possible when a voiceless segment is pronounced voiced influenced by an adjacent voiced segment. This effect occurs frequently and it is not possible to determine in every case which segment will be decisive. In examples 11. and 12. two rules may apply and the concerning alveolar plosives can be transcribed correctly whether they are produced voiced or voiceless.

5.2. Elision

13. 1ftl>fl — <freundschaftlich> (friendly)
    -Sa**ftl**IC → -Sa**fl**IC
14. 1g@l>gl — <Igel> (hedgehog)
    Qi:**g@l** → Qi:**gl**
15. 1b@n>bn — <haben> (have)
    ha:**b@n** → ha:**bn**
16. 1xm>m — <Nachmittag> (afternoon)
    na:**xm**Ita:k → na:**m**Ita:k

It is often observed that in an utterance of a word a segment is not realized although it is expected when the word is produced in isolation i.e. in its canonic form. This phenomenon is called elision. There are typical elisions that occur frequently and regularly as e.g. the elision of the apical plosive /t/ (see 13.) or the elision of /@/ in a word final syllable (see 14. and 15.). Not as frequent is the occasionally observed elision of a back fricative as /C/, /x/ or /h/ (see 16.).

5.3. Substitution of the Glottal Stop for Plosives

17. 1tm>Qm <mitmachen> (to take part)
mI**tm**ax@n → mi**Qm**axN

18. 1k#m>Q#m <guck mal> (look)
kU**k#m**a:l → kU**Q#m**a:l

Dealing with segmentation and transcription of connected speech we noticed that sometimes speakers presumably glottalize the voiceless plosives /p/, /t/, /k/, i.e. they produce them with closed glottis or substitute a glottal stop for them. It occurs after a vowel when a voiceless plosive and at least another consonant follows. This phenomenon is well known e.g. for British and American English, Swedish or Danish where a glottal reinforcement and a glottal replacement of plosives are described (see e.g. Roach [6]).

5.4. Vocalization of /l/

19. 1lz>@z <also> (well)
Qalzo: → Qa@zo:

20. 1lS>@S <falsch> (wrong)
falS → fa@S

Instead of the alveolar lateral in connected speech a vowel can often be observed. Its quality depends on the context but it is usually a rather lax central vowel and therefore described in the rule with the symbol '@'.

5.5. Consonant Epenthesis

21. 1mC>mpC <Lämmchen> (little lamb)
lEmC@n → lEmpC@n

22. 1mpC>mC <Lämpchen>(little lamp)
lEmpC@n → lEmC@n

23. 1nf>ntf <fünf> (five)
fYnf → fYntf

In consonant combinations that consist of a nasal/lateral and a following fricative/plosive an epenthetic consonant can occur. It is either an alveolar plosive after /l/ or a plosive with the same place of articulation as the preceding nasal. Respectively, this consonant can be omitted if the intended articulation is a combination of nasal/lateral and following plosive and fricative/plosive. This phenomenon can be explained with the different coordination of the articulators that are involved.

5.6. Lenition

24. 1t@n>dn <guten> (good)
gu:t@n → gu:dn

25. 1Cs>js <nächster> (next)
nECst6 → nEjst6

The term lenition refers to changes of one or more segments that are due to a higher speech rate and a not very precise articulation and reduced articulatory effort i.e. less muscular effort and energy. Voiceless fricatives may be realized voiced or in a further reduction stage as frictionless approximants. Fortis-plosives may be unaspirated, lenis plosives or even fricatives.

5.7. Realizations of /r/

The /r/ phoneme in German can be produced in many different ways, mainly depending on dialect and speaking style. But also in standard German a number of variants exist. They can be divided into consonantal (/X/) and vocalic (/@/, /6/) realizations. Here the distribution is to a large extent complementary. But in some contexts both variants are possible, most likely when the /r/ occurs at the end of a syllable after a lax vowel (see 26.). With the rules we can grasp this distribution.

26. 1!vr>!v6 <Irrtum> (error), <Hirsch> (deer)
QIrtu:m → QI6tu:m,
hIrS → hI6S

6. APPLICATION OF THE RULES

For effective further processing the canonic forms of the words that are contained in an utterance, which are taken from a lexicon, are held in form of a linear graph which is extended by applying the rules. The result is a complex graph which represents all hypothesized pronunciation variants of the concerning utterance including word boundary effects. For details about the application procedure refer to Kipp et al. [2].

7. REFERENCES

1. Barry, W. J, Fourcin, A. J.: "Levels of Labelling", Computer, Speech and Language 6, 1992, pp. 1-14.
2. Kipp, A., Wesenick, M.-B., Schiel F.: "Automatic Detection and Segmentation of Pronunciation Variants in German Speech Corpora", these Proc. of ICSLP 1996, Philadelphia.
3. Kohler, K. J.: "Segmental Reduction in Connected Speech in German", in: W. J. Hardcastle, A. Marchal (eds.): *Speech Production and Speech Modelling.* Kluwer, Dordrecht: 1990, pp. 69 - 92.
4. Lindblom, B. "Explaining phonetic variation" in: W. J. Hardcastle, A. Marchal (eds.): *Speech Production and Speech Modelling*, Kluwer: Dordrecht, 1990, pp. 403-439.
5. Pompino-Marschall, B. (ed.): "PHONDAT. Verbundvorhaben zum Aufbau einer Sprachsignaldatenbank für gesprochenes Deutsch", Forschungsbericht des IPSK (FIPKM) 30, 99 - 128, 1992.
6. Roach, P.J.: "Laryngeal-oral Coarticulation in Glottalized English Plosives" JIPA 9, 1979, pp. 2-6.
7. Wesenick, M.-B., Kipp, A. "Estimating the Quality of Phonetic Transcriptions and Segmentations of Speech Signals", these Proc. of ICSLP 1996, Philadelphia.