AUTOMATIC PHONETIC TRANSCRIPTION OF NON−PROMPTED SPEECH

Florian Schiel
schiel@phonetik.uni−muenchen.de
Department of Phonetics, University of Munich, Germany

ABSTRACT

A reliable method for automatic phonetic transcription of non−prompted German speech has been developed at the Department of Phonetics, University of Munich. This "Munich Automatic Segmentation" (MAUS) system labels and segments the phonetic constituents of spoken German in a manner similar to highly trained phoneticians. MAUS has been used to train automatic speech recognition (ASR) systems as well as to provide detailed statistical analyses of spontaneous speech (using the Verbmobil I and RVG I corpora). The MAUS system is a reliable, automatic means of testing linguistic hypotheses concerning the phonetic properties of spontaneous speech and should therefore play an important role in providing the sort of empirical data required to develop more realistic models of spoken language.

1. INTRODUCTION

In many cases our scientific work with recorded non−prompted or even spontaneous German during the last 5 years ended in results that often differ from our text book knowledge of German phonetics. In the light of these observations it is my opinion that the speech sciences including phonetics should follow a new way (beside the traditional ways that are of course still to be pursued!) to comply with the problem that often the scientific models of speech differ significantly from reality. Therefore, in part 2 I will give some arguments for computational methods on the basis of large purpose−independent speech corpora. To give an example of this type of work the third section gives a brief description of the 'Munich Automatic Segmentation' (MAUS) method, while the last part will give three examples where results from MAUS were used in different experiments or applications. The first example is a statistical evaluation of well known assimilation processes at word boundaries; the second and third example describe experiments to improve Automatic Speech Recognition (ASR) by exploiting the knowledge about pronunciation from the MAUS segmentation.

2. PRO COMPUTATIONAL METHODS

Traditional work in the empirical speech sciences (especially in phonetics) in most cases follows the approved 'divide−and−analyze' method. That is, a special question is raised, a hypothesis is formulated and then a data corpus is designed, collected and analyzed to verify/falsify the hypothesis. From the results of analysis (in most cases of statistical nature) conclusions are drawn about the nature of the problem. In some cases it has been observed that such conclusions/rules/laws/etc. are heavily data dependent and often experiments cannot be repeated successfully on different corpora. Again, this is an approved and perfectly normal way for the science community to verify the published results of their colleagues and can be observed in most other empirical sciences as well. However, the case of non−repeatable results seems to be more often reported in the empirical speech sciences (including phonetics) than others. There are several possible explanations for this situation; a likely one is the following: The corpus design or the collection method held some properties influenced by the a−priori knowledge of the nature of the following analysis. This does NOT mean that the corpus was designed to yield a positive outcome of the experiment on purpose. But for instance certain speech characteristics were suppressed in the corpus to follow the 'divide and analyze' method and this may have unexpected consequences for the investigated phenomena. In other words, it may be an inherent problem of non−prompted speech (vs. controlled read speech) itself caused by the huge diversity of natural speech signals. Every recorded speech signal is a unique event that can never be reproduced in the same way; and that is even more true for non−prompted speech. Trying to characterize the properties of different recordings a skilled phonetician can easily find about 40 more or less orthogonal factors that all have a more or less significant impact to the acoustical wave form, such as mean/max/min formants, mean/max/min f0, glottal pulse shape, syllable rhythm, nasal airflow, volume, focus characteristics, place of articulation for different phonemes, etc. If the above is true, then it follows that the empirical speech sciences should deal with large independently created corpora that in turn should reflect reality in terms of real life situations as nearly as possible.

Fortunately, this is nothing new: During the last years, following the free availability of large corpora with non−prompted or application oriented speech many of the phonetic science community have already shifted to these data. Examples are Pat Keating working on switchboard (e.g. [1]), Steven Greenberg investigating syllables in switchboard (e.g. [2]), IPO working with the Dutch polyphone database (to appear) and ongoing work in Germany on the Verbmobil corpus ([3], [4]) or the RVG corpus ([5], [6]). Interestingly enough many of these investigations have a falsifying character, that is models of fluent speech were shown to be inadequate.

To summarize this part: My argument is that phonetics (as well as other empirical speech sciences) should shift as much as possible to data that

- allow reliable statistical results
- contain 'realistic speech' (as good as it gets)
- are freely available, so that other colleagues may repeat experiments
- are not designed for one specific investigation (at best are produced by another institution).

To give one example for the possible usage of computational methods in the speech sciences the following part will give a very short outline about the MAUS method developed in Munich.

3. THE MAUS METHOD

3.1. Principle

The 'Munich AUtomatic Segmentation' (MAUS) system developed at our lab enables us for the first time to phonologically transcribe large amounts of non−prompted speech for statistical analysis. In one sentence, MAUS uses a constraint search space derived from the canonical pronunciation of the given utterance in a standard Viterbi alignment to come up with a broad phonetic transcript (SAM Phonetic Alphabet, see e.g. [8]) and a segmentation of the speech wave form. Although the segmentation still lacks the accuracy of a manual segmentation and labeling, the transcripts are within the range of the performance of skilled phoneticians on the same task. For more details about the MAUS method refer to [9], [10] and [11].

The step−by−step procedure to analyze a spoken utterance can be summarized as follows (refer to figure 1):

Input to MAUS is the speech wave and some orthographic form of the spoken text. The text may be optionally, but not necessarily extended by noise and silence markers. The text is parsed into a chain of single words (punctuation marks are stripped) and passed to a text−to−phoneme algorithm, which is either rule−based or a combination of lexicon lookup and fallback to the rule−based system. The resulting string consisting of phonemic SAM−PA symbols ([8]) is enlarged by optional inter−word silence symbols and passed to the next stage called WORDVAR.

WORDVAR is a production system with re−write rules that has expertise about German pronunciation. It takes the linear chain of SAM−PA symbols (the so called canonical pronunciation of the utterance) and computes an acyclic directed graph that represents all probable pronunciation variants of this utterance together with the predictor probability (nodes contain phonemic/allophonic symbols, while arcs represent transitions from one symbol to the next; see figure 2 for an example). Each path through this graph represents a unique possible pronunciation, while the product of all probabilities along the arcs gives the total predictor probability of this variant ([10]).

The graph and the speech wave is passed to a standard Viterbi alignment procedure that computes the best combined probability of acoustical score and predictor probability, in other words, finds the most likely path through the graph. The outcome of the alignment process is a transcript in SAM−PA together with a segmentation of the speech wave in 10msec increments.

3.2 The rule sets

MAUS can be used in two different modes using two different sets of phonological rules for the creation of the constraining pronunciation graph. A rule has the general form of

LBR > LNR ; P

where *L, B, R* and *N* are sequences of SAM−PA symbols and *P* denotes the negative logarithm of the rule probability. *L* and *R* define the left and right context of the rule; *B* is replaced by *N*.

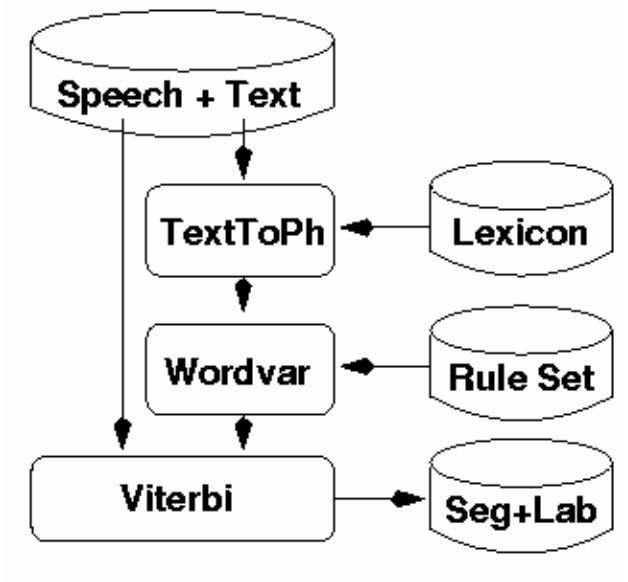

Figure 1. Processing in MAUS

The term *phonological rule is* somewhat misleading here because the underlying alphabet SAM−PA is not a pure phonological set but contains allophones as well. However, to avoid the term *phonetic−phonologic* the rules are referred to as *phonologic* throughout this paper.

P denotes the a−posteriori probability −log[*P(LNR|LBR)*] that the body *B* is replaced by *N* whenever the sequence *LBR* occurs in the canonical string of the utterance. This implies a non−recursive application of the rule set; thus the output of a rule cannot be input to another (or the same) rule. By this we gain a better control of what the rule sets possibly produce and reduce the amount of irrelevant hypotheses by several magnitudes per utterance.

Currently MAUS uses two types of rule sets (a third is under development and presented in a separate paper at this conference). In the *rule−based mode* it exploits knowledge about pronunciation variation found in the literature and empirical studies about manually segmented speech compiled into a set of approx. 6500 re−write rules. Since no statistical knowledge can be derived from literature, the rule probabilities of this set are consequently set to 1.0 ([7]). In the *statistical mode* the rule set is automatically derived from a small sample of manually segmented and labeled data (typically 1h of speech) and each re−write rule is associated with an a−posteriori probability *P* computed from the pruned observation frequency ([11]).

Figure 3 shows a result from a MAUS segmentation of the German word 'neunzigste'. As may be seen, the affricate /ts/ was reduced to an /s/ and the the second plosive /t/ has a wrong left boundary. These are typical errors that should be corrected by the still missing third last stage of MAUS.

A more detailed evaluation of the MAUS performance compared to human transcribers can be found in [11].

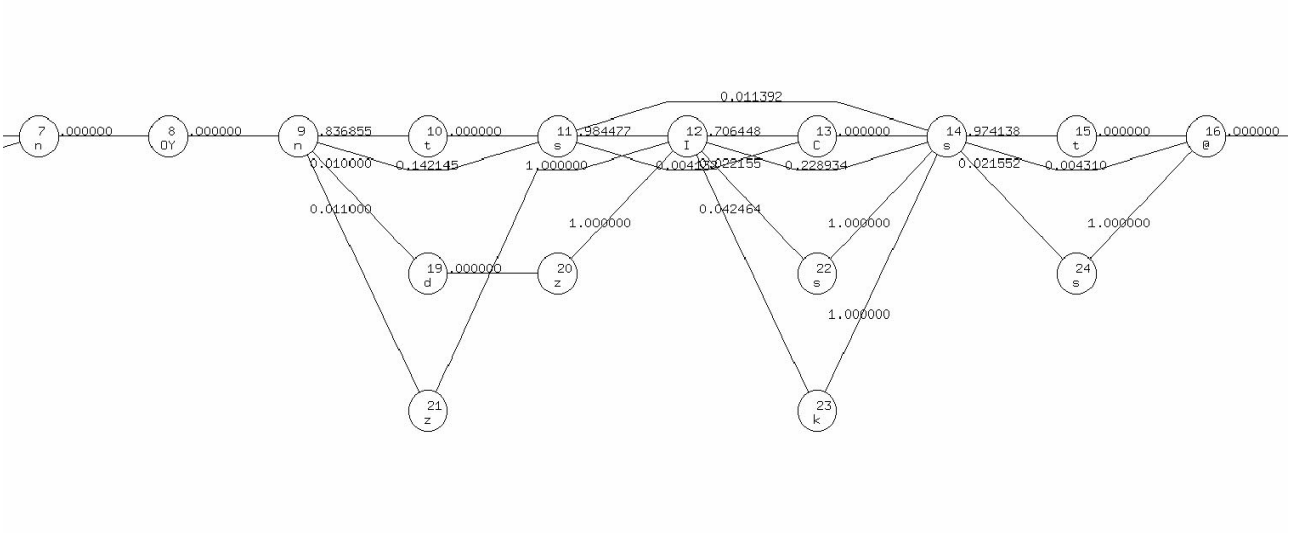

Figure 2. The MAUS pronunciation graph of the German word 'neunzigste' (ninetieth)

4. EXAMPLES

4.1. Cross−word assimilation in Verbmobil I

Aside from other investigations the simplest analysis was to verify whether phonetic/phonologic effects in non−prompted speech can be statistically verified in the Verbmobil 1 corpus as well. Our special interest was assimilation effects across word boundaries.

The VM1 corpus is a collection of 1956 dialogue recordings of 779 different speakers produced in 1993 − 1996. It contains 13910 utterances (turns) each with 22.8 words in average. In each dialog both speakers had to negotiate up to 7 business appointments using defined calendars. Overlapping speech was prevented by use of a push−to−talk−button. The whole corpus was transcribed into an orthographic markup language containing the spoken words and tags for technical noises, articulatory noises, linguistic effects such as repair, repeat, pronunciation variants, proper names, numbers, spellings and others. Parts of the corpus were labeled and segmented manually with regards to phonemic segments, prosody and dialog acts ([3]).

We choose the following re−write rules (formulated in extended German SAM−PA; '#' denotes a word boundary) from an earlier investigation done at our lab [7]:

Regressive assimilation of place of articulation

p#k −> #k	(0/95)	p(A|C) = 0
t#p −> #p	(106/196)	p(A|C) = 0.3509
t#m −> p#m	(0/2290)	p(A|C) = 0
n#p −> m#p	(16/544)	p(A|C) = 0.0286

Regressive assimilation of manner of articulation

t#z −> #s	(0/1376)	p(A|C) = 0

Progressive assimilation of manner of articulation

n#d −> #n	(360/7444)	p(A|C) = 0.0461
m#b −> #m	(0/808)	p(A|C) = 0
s#d −> #s	(7/3131)	p(A|C) = 0.0022

Voicing assimilation

t#v −> d#v	(1/1833)	p(A|C) = 0.0005
t#v −> t#f	(0/1834)	p(A|C) = 0
t#d −> #d	(3053/2404)	p(A|C) = 0.5595
t#d −> #t	(0/5457)	p(A|C) = 0

Deletion of voiceless fricatives

t#h −> t#	(41/852)	p(A|C) = 0.0459
N#h −> N#	(4/45)	p(A|C) = 0.0816
C#h −> C#	(51/947)	p(A|C) = 0.0511
x#t −> #t	(1/324)	p(A|C) = 0.0031
x#h −> #h	(3/174)	p(A|C) = 0.0169

We analyzed 303446 spoken words with the MAUS method and counted the appearance of the above assimilations. The total numbers (appeared/not appeared) are given in brackets in the second column; the a−posteriori probabilities for the occurrence of the assimilation A given the context C is given in the third column. As you can see from the raw results, six out of seventeen assimilations never occurred in the MAUS segmentation.

This result does not automatically imply that none of the listed assimilations can be found in the data; it means that the underlying statistical models in MAUS decided to model the speech wave in a way to maximize the overall Likelihood between data and model. For instance, it might be that the regressive assimilation *t#m −> p#m* was very sparse in the bootstrap set of MAUS and was therefore pruned in favor of other observations.

4.2. Regional Variation in Verbmobil I

Another experiment derived from our massive data approach was the investigation whether the knowledge of the dialect class of an unknown speaker might be a benefit for ASR in the Verbmobil speech recognizer. This work was done by N.

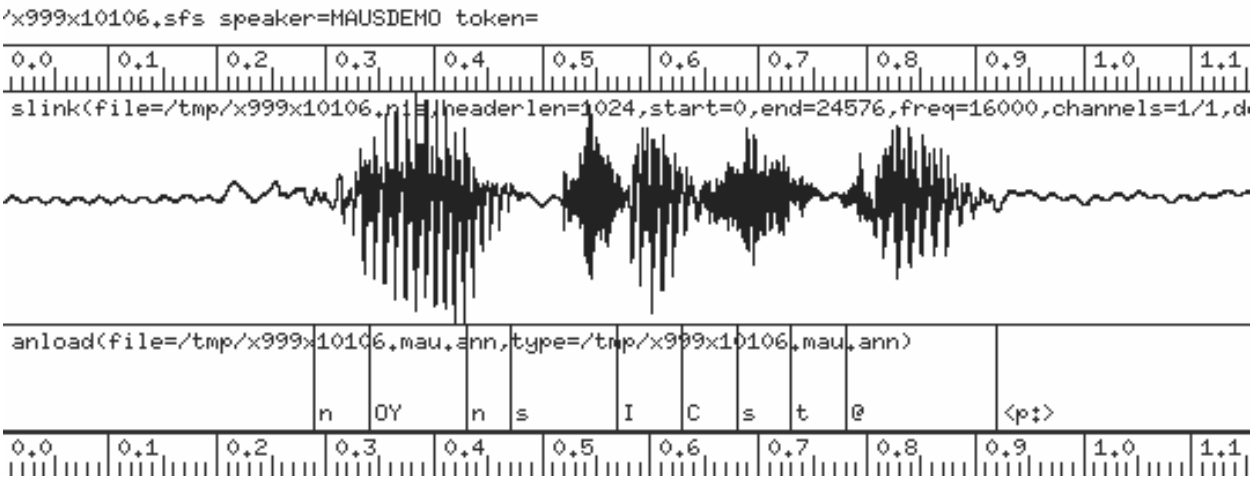

Figure 3. MAUS segmentation of the German word 'neunzigste' (Display by SFS)

Beringer and P. Regel at the Daimler Benz research center in Ulm, Germany in 1997 (details in [4]). The basic idea was to sort training speakers of the Verbmobil corpus into different dialect bins and derive dialect specific lexica for each group from the MAUS segmentation of the training corpus. In a cheating experiment where the dialectal class of each test speaker was known to the recognizer, its performance was evaluated on a bench mark task.

Although no significant improvements could be achieved in this experiment, we will continue to pursue this topic in future work at our lab in Munich on the RVG1 corpus, which contains a better controlled regional variation of speakers ([5]). Furthermore, we learned from these experiments that it is crucial to involve the acoustic modeling into the task of pronunciation modeling on the lexical level.

4.3. General Pronunciation Model for ASR

In a different approach a general statistical pronunciation model for each lexical entry was derived from the MAUS transcriptions ([12]). A standard HTK recognizer was used on the 1994 Verbmobil evaluation data to verify the model. To summarize the results the only significant improvements in terms of word recognition were achieved by using the same MAUS transcriptions for the training of the acoustical models as well as the lexical model.

ACKNOWLEDGMENTS

Andreas Kipp, Maria−Barbara Wesenick and Nicole Beringer contributed significantly to the MAUS system. Also, I would like to thank Steve Greenberg and the staff of the Realization Group at ICSI, Berkeley for their valuable comments and help.

REFERENCES

[1] Patricia A. Keating (1997): Word−level phonetic variation in large speech corpora. in: Proceedings of the Conference on "The Word as a Phonetic Unit", Berlin 22. − 23. October 1997, to appear.

[2] Steven Greenberg (1998): Speaking in shorthand − a syllable−centric perspective for understanding pronunciation variation. in: Proceedings of the ESCA workshop "Modeling Pronunciation Variation for Automatic Speech Recognition", Rolduc, Netherlands, 4−6th of May 1998, pp. 47−56.

[3] Florian Schiel (1998): Speech and Speech−Related Resources at BAS. in: Proceedings of the FIRST INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION 1998, Granada, Spain, pp. 343−349.

[4] Nicole Beringer, Florian Schiel, Peter Regel−Brietzmann (1998): German Regional Variants − A Problem for Automatic Speech Recognition?; in: Proceedings of the ICSLP 1998. Sydney, Vol. 2, pp. 85−88, Dec. 1998.

[5] Susanne Burger, Florian Schiel (1998): RVG 1 − A Database for Regional Variants of Contemporary German. in: Proceedings of the FIRST INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION 1998, Granada, Spain, pp. 1083−1087.

[6] Susanne Burger, Daniela Oppermann (1998): The Impact of Regional Variety upon Specific Word Categories in Spontaneous German. in: Proceedings of the ICSLP 1998, Dec 1998, Sydney, ???

[7] Maria−Barbara Wesenick (1996): Automatic Generation of German Pronunciation Variants; in: Proceedings of the ICSLP 1996. Philadelphia, pp. 125−128, Oct 1996.

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

[9] Maria−Barbara Wesenick, Florian Schiel (1994): Applying Speech Verification to a Large Data Base of German to obtain a Statistical Survey about Rules of Pronunciation, Proceedings of ICSLP 1994, pp. 279 − 282, Yokohama.

[10] Andreas Kipp, Maria−Barbara Wesenick, Florian Schiel (1996): Automatic Detection and Segmentation of Pronunciation Variants in German Speech Corpora. in: Proceedings of the ICSLP 1996, Oct 1996, Philadelphia, pp. 106−109.

[11] Andreas Kipp, Maria−Barbara Wesenick, Florian Schiel (1997): Pronunciation Modeling Applied to Automatic Segmentation of Spontaneous Speech. in: Proceedings of the EUROSPEECH, Sept 1997, Rhodos, Greece, pp. 1023−1026.

[12] Florian Schiel, Andreas Kipp, Hans G. Tillmann (1998): Statistical Modeling of Pronunciation: It's not the Model, it's the data; Proceedings of the ESCA Tutorial and Research Workshop on 'Modeling Pronunciation Variation for Automatic Speech Recognition', May 1998, Kerkrade/Netherlands.