

THE AT&T GERMAN TEXT-TO-SPEECH SYSTEM: REALISTIC LINGUISTIC DESCRIPTION

Matthias Jilka and Ann K. Syrdal

AT&T Labs – Research and AT&T Natural Voices
Florham Park, NJ, USA

jilka@research.att.com, syrdal@research.att.com

ABSTRACT

Like many current TTS systems the AT&T German text-to-speech system is based on the methods of unit selection and concatenative synthesis [1]. This paper highlights efforts to improve TTS quality by closely matching the speakers' original productions with linguistic descriptions. On the segmental level this is achieved by adjusting the speakers' individual productions to an established, general norm via strict monitoring and correspondingly by having the linguistic representations that control automatic alignment and TTS output, i.e. the recognition dictionary and letter-to-sound rules, reflect those original productions. The chosen standard represents a realistic form of spoken German, avoiding overly formal pronunciations. A perceptual comparison with a more traditional interpretation of German pronunciation demonstrates the positive effect of these measures on overall synthesis quality.

1. INTRODUCTION

This paper introduces aspects of the German edition (one female, one male voice) of AT&T Labs' Natural Voices (NV) multilingual text-to-speech systems. Like many commercial systems today, it relies on unit selection and concatenative synthesis.

Since the general concepts of how such TTS systems are built and function are well-known, this presentation will give a relatively brief overview of this particular TTS system's structure. However, many of the specifically German linguistic features of the system will be described, as it is a major objective of this paper to show the importance of a precise matching of speakers' original productions with the corresponding linguistic descriptions in order to achieve a high-quality speech output. The main section of this article thus provides a detailed discussion of specific features of German phonetics where such an adjustment is relevant. The effectiveness of the corresponding measures in linguistic description and speaker monitoring is tested and confirmed in a perceptual evaluation.

2. SYNTHESIS PROCESS

The German TTS system follows the general structural pattern of the AT&T Labs Natural Voices' TTS systems [2], i. e. it is based on unit selection and concatenative synthesis.

After a perceptual evaluation of natural speech recordings and synthesized test sentences, the most suitable speakers are chosen to provide an inventory of speech sounds by reading a

wide selection of texts that cover a variety of subjects and styles. Recordings are processed such that acoustic patterns can be matched to the corresponding phonetic transcriptions in a process of automatic alignment for speech recognition [3]. This provides the database from which appropriate units are selected at run-time during the synthesis process. The smallest possible units in the database are so-called half-phones, which are either the first or second half of a phone. Half-phones may be combined into phone, diphone, or larger units of speech [4]. The main principle of unit selection is to choose a path of those units with the lowest target and concatenation costs (as determined by cost function estimates of perceptual distances) which ensures that segmental and (predicted) prosodic environment are taken into consideration [5]. The necessary linguistic information for selection is provided by a text-processing module in which the input text is preprocessed (characters and complex sequences of symbols such as numbers are interpreted as text) and converted into a suitable linguistic transcription. This is achieved by letter-to-sound rules which encompass phonetic and phonological rules, morphological and syllable structure, stress assignment and exception dictionaries. In the final step of the synthesis process the selected units are concatenated.

3. AGREEMENT OF SPEECH PRODUCTION AND LINGUISTIC DESCRIPTION

Although the general system structure exerts a strong influence on the overall TTS quality of each language, linguistic parameters that require language-specific solutions are of course just as important. Accordingly, it is a major objective to attain a high agreement between the acoustic database and its corresponding linguistic description.

In order to achieve high agreement, there are three aspects that need to be controlled on the segmental level: (1) the speaker's actual productions, (2) the phonetic transcriptions provided for the recognition training (and thus the accuracy of the database of units) and (3) the letter-to-sound conversion that determines the output of the system. It is highly desirable that the latter two aspects provide identical transcriptions of the individual entries, even though the recognition dictionary may need to account for pronunciation variants that are not predictable by phonological rules. In any case, having a high degree of control over the speaker's productions constitutes a very helpful prerequisite to obtaining accurate transcriptions, especially as far as letter-to-sound rules - which are by definition of an inherently more general nature - are concerned. For this reason it makes sense to predetermine a pronunciation norm which the speaker should follow as closely as possible.

Such a procedure can be interpreted as an effort to "generalize" speech production in order to ensure precise transcriptions.

The chosen norm should represent a realistic, widely acceptable and not overly formal manifestation of spoken language, in this case of standard, so-called High German. Extra caution should be taken to cover and make the speaker aware of all important phenomena such as vowel reductions, exaggerated formality associated with read speech or popular dialect-influenced variants. These aspects can cause unwanted deviations from the norm which are easy to miss during recordings, even more so if a thorough monitoring of the speaker's productions is not possible. Some key examples for German are presented in section 3.2.

Linguistic descriptions are given which show how these phonological phenomena are transcribed. The transcriptions are meant to reflect each speaker's individual pronunciations. However, the generalizing influence exerted on the speaker's productions should to a considerable degree counteract the need for a large number of speaker- or even instance-specific transcriptions. For training the recognizer, it may nevertheless be required to include unique pronunciations in the recognition dictionary, something that is unnecessary for the TTS output dictionary/rules, where one correct transcription is sufficient.

In summary, from the point of view of speech production, a high degree of generality must be aimed for, whereas linguistic description must have the objective of reflecting each individual instance of a word's pronunciation accurately. These two elements should lead to a high correspondence between acoustic patterns and symbolic representations. While it seems obvious that this correspondence should be a major criterion for a high-quality TTS system, it must be pointed out that this factor is often greatly underestimated. Without guidance even professional speakers make numerous errors and are often inconsistent; also dictionaries usually contain many errors, and letter-to-sound rules are often too formal and general. In other words, there is considerable potential for improvement.

3.1. Recordings and linguistic descriptions

For the AT&T Labs Natural Voices German TTS system one female and one male voice were recorded. All recordings were monitored with respect to the agreed upon norm and with particular attention to potentially frequent, unwanted deviations (see section 3.2.)

For the recognition training a dictionary with ca. 400,000 entries was used, the result of merging a public domain German dictionary and German letter-to-sound rules. Extensive manual corrections were made in the newly created dictionary to ensure the correct transcription of at least all recorded words.

For TTS output, attention was focused in the areas of text normalization, preprocessing and the actual letter-to-sound rules, allowing the reduction of the dictionary to approximately 64,000 entries not (yet) covered by the rules.

Concrete examples of common problems in text processing and letter-to-sound rules as well as phenomena requiring control of both production and description are given in the following section.

3.2. Relevant potential problems in production, linguistic description and text interpretation

This section lists concrete phenomena that were addressed in the development of the German front-end from an existing rule framework and that constitute relevant potential problems for any German system. There are two major types of such phenomena: those that affect exclusively linguistic description and those which require an interaction of description and production.

3.2.1. Important aspects of linguistic description

One obvious requirement of major importance in the area of text normalization is to account for the replacement of the so-called umlauts "ä" (/E/ or /E:/), "ö" (/9/ or /2:/) and "ü" (/Y/ or /y:/) by the digraphs "ae", "oe" and "ue" in some texts, without neglecting exceptions such as "Poet" or "aufzuessen", in which two separate vowels are pronounced.¹

Also, small-scale errors in the letter-to-phoneme conversion rules have to be avoided, such as, e.g., the pronunciation of the letter sequence "sph" at the beginning of a morpheme (e.g. "Sphinx") as /sf/ and not /Sph/ due to lack of an exception to a more common rule that would pronounce morpheme-initial "s" as /S/ before "p".

The phoneme /w/ should be replaced by short /U/ in words with "u" preceding vowels (e.g., "Pinguin", "individuell") and be restricted to English loan words such as "Whiskey".

Another important measure was the addition of five phonemes (/eI/, /OU/, /D/, /T/ and /rR/) to account for the non-negligible number of English loan words in modern German. The /rR/ symbol was created to distinguish the English r-sound from its German equivalent.

3.2.2. Important aspects of spoken standard German

There are several phonetic phenomena in German that involve a high degree of variation of which the vast majority of speakers are unaware. For a TTS system it is however necessary that individual instances of such phenomena are pronounced as consistently as possible. Therefore it is quite advantageous to set a standard that is followed both by speaker and transcription norm as discussed earlier.

One of the most important issues is the treatment of schwa-elisions in morphemes ending in "en" or "el", creating syllabic /n/ or /l/. Retaining the schwa would be unrealistically formal, deleting all of them, on the other hand, would also be problematic in words with sequences of schwas like "bleibenden".

For this reason schwa elisions are handled the following way:

- when schwa is preceded by /rR/ (English r-sound), /R/, /l/, /w/ or /j/, it is not elided at all, e.g. "fahren" (/fa:R@n/), "Fohlen" (/fo:l@n/), eng. "barrel" (/bE:rR@l/).
- in sequences of schwas only the first schwa is elided (e.g. "bleibenden" = /blaIbmd@n/)

¹ Transcriptions throughout this paper use SAMPA [6]

- if a schwa preceding /n/ is elided, /n/ becomes syllabic and may assimilate to the preceding consonant, becoming syllabic /m/ (if following /b/ or /p/) or /N/ (after /k/ or /g/).

A similar case where a less formal pronunciation leads to better-sounding synthesis is the lowering of schwa to a so-called "vocalic r" (or "a-schwa" like in "Lehrer") when the schwa precedes an "r" in the following syllable. This concerns cases such as "wenigere", "Kellnerin" or "Schweineerei". This rule obviously must not apply to the prefixes "be" and "ge", e.g. in "bereit" or "gereizt").

Another important measure improving synthesis quality is the consistent transcription of vowels with non-primary stress as lax/short. This affects especially non-native words like "Ökonomie" or "Periode", which are often transcribed as tense but short in pronunciation dictionaries like the Duden Aussprachewörterbuch [7]. Syllables with secondary stress in compound words are of course excluded from this process (e.g. "Monatsmiete").

A final aspect is the influence of dialectal features. Even though in Standard German there is a three-way opposition between /E/ (short, lax), /e:/ (long, tense) and /E:/ (long, lax), the northern tendency toward replacing /E:/ by /e:/ such that words like "Käse" are pronounced /ke:z@/ instead of /kE:z@/ has become quite influential and leads to inconsistent pronunciations even by professional speakers. In the recordings and transcriptions for this TTS system, extra attention was thus paid to the correct pronunciation of /E:/.

4. PERCEPTUAL EVALUATION

In order to assess the effectiveness of the modifications described in section 3., a web-based listening evaluation of two different versions of the German TTS was performed. One version included the modifications described above, the other didn't. There were two parts to the evaluation, the first part was an A/B Paired Comparison Test, and the second part was a subjective rating test to provide overall quality ratings. Another evaluation that included ratings of a competitive German TTS system is discussed in [8].

4.1. Text Materials

Sixteen test utterances were synthesized using each of two versions of the AT&T NV German TTS female voice: one with the described adjustments to spoken German and one with the original basic rule framework. The majority of the stimulus sentences differed in only one of the aspects listed in section 3 and there were usually two test sentences per adjustment:

- /6/ for /@/ before /R/ as in "Kellnerin" ("vocR")
- tense unstressed vowels becoming lax as in "Problem" ("tu")
- lax /U/ instead of /w/ as in "aktuell" ("Uw")
- umlaut pronunciation of "ä" as /E:/, not /e:/ as in "Gerät" ("uml")
- the use of English phonemes in loan words like /eI/ for /e:/ in "Play" ("eng").
- schwa elision rules as in "bleibenden" ("se")

Only one stimulus sentence using the schwa elision rules exclusively was employed, but schwa elisions occurred numerous times in those stimulus sentences that tested the cumulative effect of the listed pronunciation modifications:

- combination of "tu" and "se" ("tuse")
- cumulative effect in stimuli with several instances of "se", "tu", "uml", "eng" and "vocR" ("cum")
- stimuli randomly selected from German on-line news, included instances of "se", "tu", "uml" and "eng" ("rdm")

4.2. Listeners and Procedure

15 adult native speakers of German volunteered to participate. All were employees of AT&T or family members of employees.

Listeners were instructed in German (on the web-site) to listen to the audio files by clicking on the icons. All subjects were unaware of the identity of the pronunciation rules associated with each file, and the order of files was random. After listening to the two versions of a file, they were instructed to choose the file they believed sounded best. In the subjective ratings test they were asked to rate the overall speech quality on a scale from 1 (Schlecht/Bad) to 5 (Sehr Gut/Excellent). Participants were told that they may listen to the files in any order, and as many times as they liked. They were instructed to listen with headphones.

4.3. Results

4.3.1. Preference Test

In the preference test the modified pronunciations were favored 78.3% of the time over the original pronunciations, and each individual listener showed an overall preference for the modified pronunciations. Results of a one-sample t-Test (two-tailed, testing $\mu = 8$, the number of preferences for each TTS version expected by chance alone) indicated that listeners' preference for the modified pronunciations was statistically significantly higher than would be expected by chance ($t = 8.962$, $df = 14$, $p < 0.0001$). Column 2 of Table 1 shows the group results for the individual test sentences.

4.3.2. Ratings Test

In the ratings test, on average, the modified pronunciations were rated 0.63 MOS higher than the original ones. A repeated measures ANOVA was performed, with TTS version (2) and utterance (16) the within-subject factors in a fully factorial design. There were significant main effects for TTS version ($F(1,13)=60.848$, $p<0.0001$), indicating that the ratings for the modifications were statistically reliably higher, and for stimulus ($F(15,195)=20.065$, $p<0.0001$), reflecting significant rating differences among the 16 test utterances. There was also a significant version by stimulus interaction ($F(15,195)=8.321$, $p<0.0001$), indicating that rating differences between the modified and original versions differed significantly depending upon the test utterance.

Ratings by TTS version and stimulus are listed in Table 1 (column 3 shows ratings for the modified pronunciations, column 4 those for the original rule framework). For 8 of the 16 utterances, ratings for the modified pronunciations were

significantly higher. For 7 stimuli, there was no significant difference between the two versions, and for one stimulus, the original pronunciation was rated significantly higher.

Stimulus	% Mods. Pref.	Mods. Rating	Orig. Rating
se	60.0	3.857	4.286*
vocR1	86.7	4.143*	3.286
vocR2	86.7	3.286*	2.643
uml1	100.0	3.786*	1.786
uml2	73.3	4.571*	4.143
eng1	60.0	3.786	3.857
eng2	40.0	4.357	4.143
tu1	53.3	3.857*	1.929
tu2	80.0	4.786*	3.714
Uw1	86.7	3.571	3.357
Uw2	100.0	4.071*	2.857
tuse	100.0	4.571	4.429
cum1	53.3	2.857	2.643
cum2	93.3	3.500*	2.286
rdm1	80.0	3.214	3.071
rdm2	100.0	3.286	2.929

Table 1. Summary of results for the individual stimuli (* indicates a significantly higher rating ($p < .05$))

4.3.3. Analysis

In both tests listeners judged the modified pronunciations to be of a higher quality significantly more often than the ones based on the original rule framework. There was reasonably good, but not perfect agreement in results for individual utterances across the two parts of the evaluation. Of the 8 utterances rated significantly higher for the modified pronunciations, the average preference in the paired comparison test was 84%. Of the 7 utterances whose ratings were statistically equivalent, the average preference for the modified pronunciations was 74%. The stimulus "se", the only one for which the original pronunciation was rated significantly higher, the paired comparison results also favored the modified version, but only by 60%.

This result suggests that in some words with multiple schwa elisions, especially those involving assimilation of /n/ to the preceding obstruent no schwa elision may be preferable for TTS quality. Similarly, the German listeners did not completely appreciate the use of English phonemes in English loan words ("eng" stimuli), e.g. /w/ for /v/ in "Whiskey" or the diphthong /eI/ for /e:/ in "play". In both cases the smaller inventory of suitable units, either syllabic /m/ and /N/ or English phonemes as opposed to the more common "regular" German phonemes could also contribute to these results.

In general, however, the modified pronunciations introduced such as the use of lax variants of unstressed vowels in non-native words or the replacement of the glide /w/ by lax /U/ before vowels (unless in relatively recent English loan words, spelled with "w") clearly improved synthesis quality.

The test sentences where these phenomena occurred more than once, either by design ("cum") or randomly ("rdm") confirm this impression.

5. SUMMARY AND CONCLUSIONS

This study provides an account of the AT&T Natural Voices German text-to-speech system's structure as well as a step-by-step description of the process of converting a given text to natural-sounding synthetic speech, underlining the centrality of unit selection and concatenative synthesis. Concrete examples are presented, focusing on aspects of the pronunciation of spoken standard German that have to be controlled in production and transcription to ensure consistency and higher synthesis quality. Contrary to the usual emphasis on aspects such as data processing or signal processing, special consideration is given to a high agreement between the speaker's productions in the acoustic inventory and the linguistic representation of these productions. It is argued that this is an essential and often underestimated factor contributing to both the accuracy of the recognition process used to build the database of speech units and the correctness of the eventual synthetic speech output.

A perceptual evaluation comparing corresponding modifications of transcription rules to a more formal-oriented rule framework confirms this both as far as preference in an A/B Comparison Test and a Ratings Test are concerned. Readers are invited to evaluate the synthesis quality themselves at <http://www.naturalvoices.com>

6. REFERENCES

- [1] Hunt, A., and Black, A. Unit selection in a concatenative speech synthesis system using a large speech database. In Proc. ICASSP-96, Vol. 1, 373-376, 1996.
- [2] Beutnagel, M., Conkie, A., Schroeter, J., Stylianou, Y., and Syrdal, A. The AT&T Next-Gen TTS system, *Joint Meeting of ASA, EAA, and DAGA*, Berlin, Germany, Paper SASCA_4, 1999.
- [3] Makashay, M. J., Wightman, C. W., Syrdal, A. K., and Conkie, A. Perceptual evaluation of automatic segmentation in text-to-speech synthesis. *Proc. ICSLP-2000*, Beijing, China, Vol. II, 431-434, 2000.
- [4] Conkie, A. Robust unit selection system for speech synthesis. *Joint Meeting of ASA, EAA, and DAGA*, Berlin, Germany, Paper 1PSCB_10, 1999.
- [5] Wightman, C. W., Syrdal, A. K., Stemmer, G., Conkie, A., and Beutnagel, M. Perceptually based automatic prosody labeling and prosodically enriched unit selection improve concatenative text-to-speech synthesis. *Proc. ICSLP-2000*, Beijing, China, Vol. II, 71-74, 2000.
- [6] Wells, J., Barry, M., Grice, M., Fourcin, A. and Gibbon, D. Standard Computer Compatible Transcription, Esprit Project 2589 (SAM). Doc. No. SAM-UCL-037, Phonetics and Linguistics Dept. UCL, London
- [7] Mangold, M. *Das Aussprachewörterbuch*, Dudenverlag, Mannheim, 1990
- [8] Strom, V. From Text to Prosody without ToBI, to appear in: *Proc. ICSLP-2002*, Denver, USA