Technical Report CSE-98-015

Rapid Prototyping of a German TTS System

8 June – 15 July 1998

Michael W. Macon, Alexander Kain, Andrew Cronk
Oregon Graduate Institute

Horst Meyer
University of Bonn

Karin Mueller, Bettina Saeuberlich
University of Stuttgart

Alan W. Black
University of Edinburgh

sponsored by *fonix* Corporation

September 23, 1998

The aim of this project was to build a German synthesis system within six weeks, using the Festival synthesizer as a starting point. Festival was developed by Alan Black and Paul Taylor at the Centre for Speech Technology Research, University of Edinburgh, UK [1].

There were three German students invited to take part in the festival summer course:

- Bettina Saeuberlich, a computer science student at the University of Stuttgart,
- Horst Meyer, a phonetics student at the University of Bonn and
- Karin Mueller, a computational linguistic student of the University of Stuttgart.

Michael Macon served as coordinator of the project, and Alan Black was a consultant who visited OGI for the first 10 days of the project. The various modules developed in the project will be described in the subsequent sections.

1 Phonetic Alphabet

There are many more German phonemes than in English - for example French and English phonemes of loanwords - they had to be into the alphabet. The following phonemes are used in our lexicon (based on the SAMPA system)

SAMPA	EXAMPLES (orthographic)		
_	(silence)		
		aI	EIns, kAIser
p	Pier, aB	OY	Aeusserung, nEU
b	Bier, KraBBe	aU	AUf, schAU
t	Tier, GraD		
d	Dir, eDel	@	sehEn, bEsagt
k	Kasse, taG	i:	Igel, bIEten
g	Gasse, eGal	I	In, bItten
		y:	Uebung, hueten
f	Vogel, schlaF	Y	Ypsilon, huetten
v	Wasser, eVentuell	e:	bEten, schnEE
s	aSt, haSS	E	bEtten, gaeste
z	Sieb, beSen	E:	aesen, geblaese
S	Spaet, aSCHe	2:	oefen, moegen
Z	Genie, DSCHungel	9	oeffnen, koennen
x	naCH, doCH	u:	bUHlen, gUt
C	diCH, honiG	U	lUstig, bUtter
h	Hut, aHorn	o:	Ofen, kOHl
		O	Offen, tOpf
pf	PFerd, toPF	a:	wAr, wAHr
ts	Zwei, plaTZ	a	An, kAnn
tS	kuTSCHe, Cello		
		E~ (= E~:)	bulletIN
m	Mut, haMMer	a~ (= a~:)	pendANT
n	Nase, kaNNe	9~ (= 9~:)	parfUM
N	eNG, baNGe	o~ (= o~:)	feuilletON
		E~	IMpair
l	Liebe, haLLe	a~	pENdant
R	Riese, kRaut	o~	nONchalant
6	opER, deR		
j	Jetzt, Jagd		

The short nasalized vowels were skipped because of the consideration that normal German speakers are not able to make a difference in length of nasalized vowels.

2 Pronunciation Lexicon

The lexicon that is used by the German synthesizer is extracted from the CELEX lexicon available from the Linguistic Data Consortium. Unfortunately, this lexicon and works derived from it cannot be redistributed under the guidelines of the LDC, so it will not be part of the free distribution of this work available on the WWW.

The lexicon had to be converted to the SAMPA phonetic alphabet and to the format that is required by the Festival system. CELEX is a full form lexicon containing approximately 360,000 entries. Each inflected word is listed as its own lexicon entry. This is just a compromise, as German is a productive language—this means it is possible to create an indefinite number of new compound words. The system should also include a morphological module that is able to split the words into morphemes, but development of a morphological parser for the language was not undertaken. However, up to this point the lexicon and letter-to-sound rules are sufficient to synthesize most of the German words.

If any word appears that is not part of the lexicon, the pronunciation will be found by letter-to-sound-rules, which were statistically trained from the CELEX corpus based on a recently developed method [2]. This gives the synthesizer the ability to generalize to pronunciations of words not in the lexicon, based on similar words it does have.

Postlexical rules were developed for glottal stop insertion – a commonly-occurring feature of German that is sometimes coded in a lexicon (as in the Hadifix system at Bonn), but can be approximated by rule. The postlexical rule inserts a very short pause after a word if the next word starts with a vowel or a diphthong. This short pause implies that the diphone synthesizer will use units that adjoin silence, and hence will contain a glottal stop.

Glottal stops also occur between morphemes in many German words. For this purpose, a simple morphemic postlexical rule was written which inserts a glottal after different prefixes like (ent-, um-, aus-, dis-, des-, un-, mis-, ver-, in-, ge-, ko-, be-, ab-) if there follows a morpheme that starts with a vowel or a diphthong.

Another problem had to be solved about the CELEX-lexicon which has only one phoneme for /x/C/, even though these are significantly different allophones. This can be disambiguated via a phonetic rule: /x/ follows a low vowel and a /C/ only can follow a high vowel.

3 Word Stress Rules

Word stress is realized by a simple rule. Every content word is accented, meaning it receives greater energy than non-content words like prepositions, articles, auxiliaries, question words, conjunctions and modals. The set of non-content words can be explicitly listed, and words not found in this list are assumed to be content words.

4 Duration Model

A duration model statistically trained from the (English) Boston University Radio News Corpus was used to produce segmental durations. Although this is obviously not the ideal method, the results are quite reasonable. Each phoneme is described by a vector of distinctive features. A statistical prediction algorithm predicts the duration of each segment in the context of the features of the segments surrounding it, as well as place in the phrase, stress, and other factors. The timing of English and German syllables is quite similar because of historical relationships between the languages, so the results are reasonable. (This technique would be less likely to work for, say, French, which is syllable-timed.)

5 Intonation Model

A simple intonation model was incorporated. The lexically-stressed syllable in each content word is given a pitch accent. Based on these pitch accents, a statistical model (again trained for English from the BU Radio Corpus) predicts the value of F_0 at the beginning, middle, and end of the syllable [3]. We compared this with other simple intonation models (e.g., steady fall across the sentence with hat-shaped accents on each stressed syllable), and found it to be most appealing for sentences and paragraphs.

There is currently no phrase breaking mechanism for points in the sentence other than at commas and end of sentences. In German there very long sentences can appear, thus the F_0 contour can fall to a very low level.

6 Text Preprocessing

The German text preprocessing was developed by Mark Breitenbuecher and Gregor Moehler at the Institute of Natural Language Processing at the University of Stuttgart. The manual of the original version can be obtained from (http://www-stud.ims.uni-stuttgart.de/ breitenb/) It was modified to work within Festival version 1.3 and some functions were expanded.

The text preprocessing takes care of the following problems: expansion of numbers, abbreviations, special characters and punctuation.

The expansion of numbers can be separated into:

- dates (in the format of “16.7.1998” or “16.7.98”),
- fractions (format: cardinal/cardinal),
- Verhaeltnisse (format: cardinal:cardinal),
- telephone numbers (format: leading zero followed by an number of cardinals and a ’/’ again followed by a number of cardinals),
- compound words starting with a number (e.g. “16jaehrig”)
- years (e.g. “1989”)
- time (format: “13:23:34” or “13.34 Uhr” or “13.34 h”)
- currencies (e.g. “$ 3,56”, “4,566 DM” “2,-” ...)
- comma numbers
- ordinal numbers (depending on different contexts, ordinal numbers are expanded with different suffixes)
- cardinals (format: just numbers or numbers in blocks of three)
- roman numbers (e.g. “Elizabeth II” is expanded to “Elizabeth die zweite”)

The expansion of abbreviation deals with:

- common abbreviations (e.g. “z.B.”) with and without punctuation
- compounds of abbreviations (e.g. “Vers.-Ges.”)
- measurements (e.g. “5 km”)
- word that contain a capital letter in the middle are spelled

Special characters and punctuation:

- Special characters are expanded, regarding the context in which they appear.
- Punctuation is different whether it is at the end of a sentence or belongs to a number or an abbreviation.

7 Waveform synthesis

The algorithms are based on concatenative synthesis, so two diphone databases were collected - one male (AXK) and one female (BCS).

7.1 Recording

The diphones were extracted from carefully designed nonsense words. All combinations were generated by and put into a reasonable carrier context. Phonotactic constrains of the German language were used to trim this list down to a set of about 2240 used in the synthesizer (e.g., “Auslautverhaertung”– voiced stops and voiced fricatives become unvoiced when they appear at a morpheme boundary).

Recordings were made on two separate days (one per voice) at a Blue Dog Recording Studio in Portland. Since frequent breaks are necessary because of fatigue of the subject being recorded, recording each voice required about 6-8 hours of studio time.

The speech was recorded on digital audio tape (DAT) to assure high quality of the data. A cue sound was played between the items in order to excise the data from DAT more easily, although several time-consuming steps were still required to transfer and preprocess the data into a manageable format. In the future, we hope to explore high-quality direct-to-hard-disk recording methods that will cut out many of these steps while keeping the audio quality high.

7.2 Labeling

The AXK (male) voice was hand-labeled. The CSLU Toolkit application `SpeechView` was used for this purpose, and took two people approximately 3 working days to complete. The BCS voice was first auto-labeled with a German speech recognizer developed at Fluent Speech Technologies. This process produced reasonable results (i.e., the speech was intelligible). Subsequent hand-optimization of the labels produced dramatic improvements, however.

7.3 Signal processing

The `OGIresLPC` synthesizer developed at OGI [4] was used as the signal processing "backend" of the system. This takes diphone waveform files, coded as residual-excited LPC parameters, and performs the necessary concatenation, pitch/duration change, and smoothing of the diphone waveforms.

8 Evaluation

The most significant evaluation process carried out was in the optimization of the BCS diphone set after forced-alignment to correct bad labels. This was performed by creating a script that generated nonsense syllables similar to the ones used in the recording. Listening to these was the most efficient way to listen to *all* of the diphones in a reasonable amount of time. When poorly labeled diphones were heard, these labels were fixed manually.

A large set of German sentences were also taken from various German web sites and used as more natural testing material to make informal judgements of needed modifications. Because of the short duration of the project, no "formal" multi-listener evaluations were conducted. However, the synthesizer has been entered into the multi-lingual TTS Evaluation Session at the 3rd ESCA/COCOSDA Speech Synthesis Workshop in Jenolan Caves Australia, to take place in December. At this workshop, several hundred expert speech scientists will listen to the output and rate the quality along several different scales.

A website will be developed under http://www.cse.ogi.edu/CSLU/tts/ in September 1998 to allow anyone in the world to test the system. This web site will also be used as the distribution point for all non-proprietary parts of the code developed – this excludes the lexicon, letter-to-sound rules, and German text preprocessing. All other parts will be distributed free for non-commercial research use. Commercial rights to the work are reserved by *fonix* Corp.

9 Summary

The synthesizer developed in this short project is of reasonable quality. In the end, much of the time during the project was spent in collecting and processing data for the concatenative synthesizer. Several problems occured in this process (corrupted data files, computer crashes, etc.), which left the team with less time than desired to work on prosodic aspects of the system. Using a less data-intensive synthesis method (i.e., formant synthesis) would probably lessen the problems associated with data collection, but also introduce a host of new ones.

Festival was an ideal framework in which to conduct this work. The only C++ code added to the system consisted of a few simple routines associated with the text preprocessing module. Otherwise, a completely new language capability was added to the system in 5 weeks, by relatively inexperienced students, and without adding new C++ code to the system. This demonstrates that Festival is able to be truly "multi-lingual" (at least for Western languages).

The participants wish to thank *fonix* Corporation for generous support of this work.

References

[1] A. W. Black and P. Taylor, "The Festival speech synthesis system: System documentation," Tech. Rep. HCRC/TR-83, Human Communication Research Centre, University of Edinburgh, Scotland, UK, January 1997. available from `www.cstr.ed.ac.uk/projects/festival.html`.

[2] V. Pagel, K. Lenzo, and A. W. Black, "Letter-to-sound rules for accented lexicon compression," in *ICSLP* (to appear), December 1998.

[3] A. W. Black and A. J. Hunt, "Generating F_0 contours from ToBI labels using linear regression," in *ICSLP*, October 1996.

[4] M. Macon, A. Cronk, J. Wouters, and A. Kain, "`OGIresLPC`: Diphone synthesiser using residual-excited linear prediction," Tech. Rep. CSE-97-007, Department of Computer Science, Oregon Graduate Institute of Science and Technology, Portland, OR, September 1997. available from `www.cse.ogi.edu/CSLU/tts`.