HOW WELL DO HUMANS RECOGNIZE SPELLINGS IN GERMAN TELEPHONE SPEECH?

Christoph Draxler
Department of Phonetics, University of Munich,Germany

ABSTRACT

We have set up a WWW-based experiment to measure the performance of humans in a spelling task. This measure can then be used as a baseline for the evaluation of automatic spelling recognizers. The speech material consists of 500 spellings of artificial words of the German SpeechDat(II) database of telephone speech. Each test person listened to 40 spellings via a WWW browser and entered the letters he or she heard into a database via a form. A total of 105 persons participated, but only some were able to complete a full session. The overall results show that 50.0% of all spellings were recognized exactly. In 37.0% letters were substituted by other letters, and in 13.0% letters were missing or inserted. For spelling by letters, the incorrect recognitions can be attributed to phonetic proximity of the letters names. Spelling by name eliminates incorrect recognitions because of phonemical proximity, but introduce new errors.

1. INTRODUCTION

In the SpeechDat(II) project, a total of 25 speech databases for the fixed and mobile telephone network and 3 speaker verification databases were collected for 21 European languages [2]. SpeechDat(II) databases contain speech material for the development of voice-operated teleservices, i.e. digits, dates, times, application words and phrases, geographical, company and person names, phonetically rich words and sentences, and read and spontaneous spellings.

The German fixed telephone network database was collected and annotated by the BAS [3] at Department of Phonetics and Speech Communication (IPSK) of the University of Munich under a subcontract with SIEMENS AG, Germany. The database contains recordings of 4000 speakers from the three environments home, office, and public phone, and the database is divided into disjoint and demographically balanced subsets for training (3500 speakers) and testing (500 speakers).

In most voice-driven teleservices spelling is an important task, e.g. in telephone directory services, information retrieval, etc. Speech recognition for spelling is a particularly difficult task because the duration of the speech signal for a letter is very short, and there is a large number of similar sounding letters – this is true not only for German, but many other languages. Also, in telephone speech, the signal quality is reduced, and the signal properties needed to distinguish letters are lost during transmission. Furthermore, for spellings, and in particular for artificial letter sequences, there is no general language model. Finally, although spelling by name is often used for disambiguation, the words to represent a letter are used in an ad-hoc, non-standard and often incorrect manner, especially for the less frequent letters.

In our experiment we attempt to measure the performance of humans for recognizing spellings of arbitrary spelling sequences. The result of the experiment may then serve as a baseline for the evaluation of automatic spelling recognizers.

The remaining paper is structured as follows: section 2 describes the setup of the experiment. Section 3 reports on the results collected. Section 4 discusses the results and section 5 contains a summary and an outlook on future work.

2. EXPERIMENT

The spelling experiment is based on the WWW to allow access from remote locations and to guarantee platform independence. Remote access is necessary to avoid a regional bias of the test persons. Platform independence is necessary to allow a large number of test persons to participate, and to avoid complex software installations on the test person's host machine.

For the experiment, a test person registers for participation and then begins with the experiment. During the experiment, the test person listens to 40 randomly selected spellings in sequence and for each spelling enters the recognized letters into a form. The input is checked for formal consistency. If it is found to be syntactically correct, it is stored in a database and the next speech signal is played.

2. .1 WWW Implementation

The WWW is a client-server system. The client, a WWW browser, e.g. Netscape Navigator or Internet Explorer, requests documents from a WWW server. The server returns the requested documents to the client (Figure 1).

HTML-formatted documents, e.g. WWW pages with forms, are displayed directly by the client. Other documents, e.g. signal files, are output via plug-ins or external helper applications installed on the client machine.

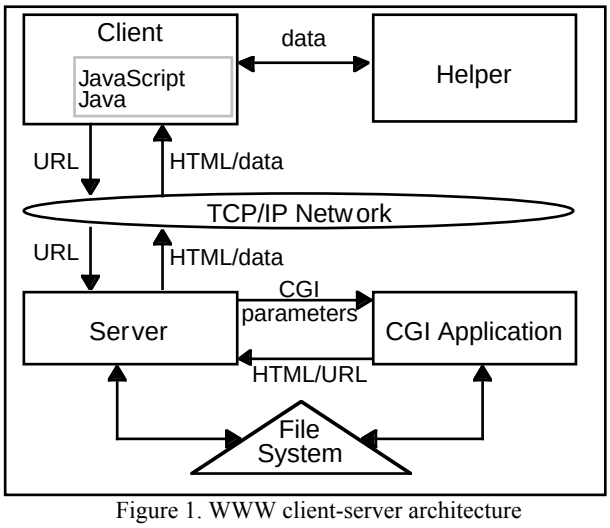

Figure 1. WWW client-server architecture

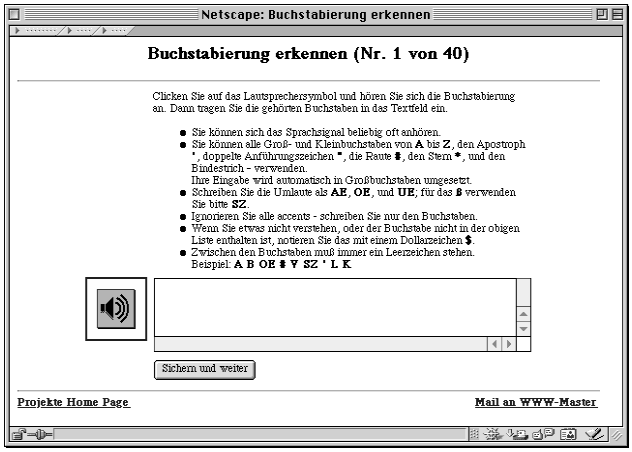

Figure 2. Spelling experiment data entry form

Client-side computations can be used to validate user input into a form before this data is sent to the server. These computations are performed by scripts embedded in the HTML pages.

In the spelling experiment, the WWW server is a standard Apache server under Linux. The selection of signal files and data storage is implemented by perl scripts from the WWWTranscribe toolbox [1] which are called via the standard common gateway interface (cgi). Signal files are formatted as WAV files, with the original alaw encoding from the SpeechDat(II) recording platform. The client-side scripts are implemented in JavaScript 1.2.

On the client machine, a standard WWW page with a form is displayed (Figure 2). This form contains instructions for data entry, a speaker symbol to start audio output, an editing panel and a button to submit the spelling text.

2. .2 Letter sequences

The test set specification of the FIXED1DE German SpeechDat(II) database contains recordings of 500 speakers. From this test set, the artificial word spelling items were selected. Note that no such artificial sequence was recorded for one call in the test set; for the experiment, the read spelling of a randomly selected word was used.

On the SpeechDat(II) prompt sheets, the spelling items consist of upper and lower case letter sequences with 8 and 10 characters each; the letters consist of the German alphabet including the umlaut letters *Ä*, *Ö*, *Ü*, the *ß*, three accented letters *é*, *è*, and *à*, the punctuation marks - and ‘, and # and * from the keyboard of modern telephone handsets. Note that the letters are distributed evenly over the prompting material for the entire database, but not necessarily over the recorded items or the test set specification.

For the spelling experiment, the allowed input was restricted to the letters *A* to *Z*, -, ‘, # and * plus *AE*, *OE*, *UE*, and *SZ* for the umlaut vowels and *ß*. Accents were not allowed. Blank spaces were required between each symbol to avoid ambiguity, and all input was automatically converted to upper case letters.

2. .3 Test person recruitment

Calls for participation in German for the experiment were published on the project information web page of the IPSK and posted internally in the department and in other departments of the University of Munich. Furthermore, speech related mailing lists were used, and finally, calls were posted to 10 national and 2 international newsgroups (*comp.speech.users* and *comp.speech research*).

The participants were offered a telephone card (value 6,-- DM, i.e. ~ 3,-- Euro) for a complete session. A complete session took between 25 and 60 minutes, depending on the download time for the signals (approx. 6.6 MB per session) and the time it took the test person to enter the data.

3. RESULTS

A total of 126 test persons registered for the experiment, but only 105 actually did take part. Of the 105 participants, only 38 were able to finish an entire session with 40 spellings, and only 18 could do more than 10 sessions. In total, 2053 spellings could be collected. Of these, 22 were discarded because they were duplicate or empty entries, so that 2031 spellings, i.e. 48.4% of 4200 expected spellings, could be used for the analyses.

3. .1 Test person demographics

The gender distribution of the test person was 85 male (80.9%), 20 female (19.9%).

3. .2 Spelling recognition results

The original SpeechDat(II) spelling transcriptions are verbatim orthographic transcriptions of the uttered speech. These transcriptions were produced by skilled transcribers. The transcriptions are of a high quality – in the final SpeechDat(II) validation, the error rate for the long items, i.e. spellings, was 5.1%. These transcriptions contain markers for noise, signal truncations, and mispronunciations. For the spelling recognition, the SpeechDat(II) transcriptions were normalized to the same format as used by the test persons: the SpeechDat(II) orthographic transcription *[spk] Gustav Friedrich Martha Theodor Anton Ludwig Y Übel Siegfried Siegfried* is normalized to *G F M T A L Y UE S S*.

In the comparison of the normalized transcription and the actual spelling three types can be distinguished:

1. Normalized transcription and spelling are identical
2. Normalized transcription and spelling have the same length, but differ in some letters
3. Normalized transcription and spelling have different lengths and/or contain different letters.

Table 1 displays the three types of spellings for both spelling by letter and spelling by name.

	Spelling by letter		Spelling by name	
Type	Count	Percent	Count	Percent
1.	888	48.3%	127	65.5%
2.	719	39.2%	32	16.5%
3.	230	12.5%	35	18.0%
	1837		194	

Table 1: Recognition results

For types 1 and 2 and spelling by letter, a total of 13560 letters were counted, 11458 (=84.5%) of which were recognized correctly. 2102 (=15.5%) letters were not recognized correctly but mistaken for some other letter. For spelling by name, the corresponding figures are 1282 letters, 1121 (=87.4%) recognized correctly, and 161 (=12.6%) recognized incorrectly. Table x contains the confusion matrix for the type 2 spellings for spelling by letter.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AE	OE	UE	SZ	*	-	#	'	$
A	577		1	6	5	1		10	2	8		2	4		3		1	1	1		2	1		2	4	5	12		3		10				
B		276	4	10	4		2									2			3	1			13					7			6				
C	1	6	528	2	2	1	2		3	4	2			1	6	5		1	1	17		6	1			9		10	2	1	2	1		1	
D		22	18	265	2		8	2	2	1					1	7			1	1		5	6			1	1	4	1			1			
E	4	5	4	7	783		10	4	38	2	3				1	2	1	6	2	6				1	1		29	3	4		6	5	5		
F		1	2	3	4	362	3	2					2	1					121	2		2		2		6	1		2			1			
G	4	1	1	8	1		457	1	8	1	6		5	5			2		4	1			1			1								2	
H	18	4		1		8		345	1		6						3	2				1	4		6		3	2				5			
I		4		9	28		4	1	510	1	3	4			1		1		1	3			1		1	9	8		19		9	1		1	
J	3		2		6		1		6	415			5		4				5	6	1		1					2			8	6			
K	1					2		5			428		1		2	3	1				7					11	5				6	1			
L					7		1	2	17	6	9	297	2	3			1			5				2		1	7		1						
M	3		9				1	8	1				377	18		6		1	9		1	3							1		4				
N	3					1	7	1	4	4		2	44	316	1									6	1				1					5	
O	3		2		4		1							2	518	1					26		1	2	5		1	7	1			1	3		
P	2	1	8	5	5		2					2			1	342				22			14		1			1			1	6			
Q					2			2		2	1						424			1	4	5		3			1					2		3	
R	1		4		5	2		1	4		1	1	7		2	1		389	1		1	1	2	1			11					8			
S			1	1	4	61	1	2				1		4			1		319			3		3		3				15	1	4			
T	10	2	2	9	7		5			1		1		5	4	32	8			252	1		3					11			3	1			
U	4	1		1	2				3		1		7		5		3	1			393					1			12				2		
V	1					4			1	6	2	5		2	1	1		2			1	402			6							4		7	
W	3	17	1	6	6	3	9			6	2			2		8	1		1	5	1	1	418		1	2	5		5		5	11			
X			6	1			5	1	2		1					1			1					380	1	2	3	3	2					
Y			6				6	4			1	1	2	4	1		2							2	522			2	15		2	15		5	
Z	2		3	1	1	2		4		2		9							1	5	2		2		6	359			3	3	9	5			
AE					11			5	2	1		12	1		1		5	1				1					432	3	1			2		3	
OE	1	3		7	5	1	5	3		3	3		1		2	6		5			1		5		5		1	237	12	6			1		
UE	7		3						6		1	1		2	2			7		6	2			1	2			2	492	2	5				
SZ									1														10						1	137					
*				1	2	1						3			5															181	1				
-											1		6					2		3		1				1					230	4	6		
#												4						1														78			
'	4	3	2		9			1	1				2			2	3	1		6				1	9	1					3		73	
$	4				7	2		1			3	7				4	4						1	1	4	1		2	4		7	3	8	5	
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AE	OE	UE	SZ	*	-	#	'	$

Table 2. Confusion matrix for SpeechDat(II) transcriptions and spellings of same length

4. DISCUSSION

The discussion covers two topics: the technical problems experienced during the experiment, and the spelling results proper.

4. .1 Technical problems

In the WWW, the provider of a service has no control over the configuration of the client accessing the server. Although there exist ISO or de facto standards for HTML, JavaScript, and audio formats, this is no guarantee that a service will function properly. A general problem is that of outdated versions: for example, the spelling experiment requires regular expressions which have been added to JavaScript only in version 1.2 – older browsers do not support this version.

A second general problem is that of software quality. Although JavaScript 1.2 is supported by both Internet Explorer and Netscape Navigator, its implementation differs in both browsers and also across operating systems. As a consequence, some test persons experienced serious system crashes while playing audio data or executing scripts.

A more specific problem is the configuration of a particular client: JavaScript may have been disabled manually by the user, or external helper applications, e.g. to play audio signals, are disabled.

In the spelling experiment, clients are requested to perform two tests to check whether they could participate: JavaScript and audio. It turned out that the first test was too liberal: if a JavaScript alert box could be displayed, it was assumed that the client understood JavaScript. However, all versions of JavaScript can display the alert, but only the most recent ones feature regular expressions. As a consequence, test persons started the experiment but could not enter data.

A further problem was download time: although the IPSK is connected to the outside world via a high-speed fibre optic link, many test persons got very low transfer rates and gave up before completing the experiment.

Quite a few test persons apparently were interested in getting only the reward: they entered meaningless data, or they registered but never entered the experiment.

Finally, the demographics of the test person population do not match that of the overall population. One reason for this may be that the demographics of the Internet users, especially those that have free access to the Internet, e.g. in their office, is different from the general population.

4. .2 Recognition results

One of the original goals of the experiment was to achieve reliable results by having a large population of test persons recognize a statistically balanced set of utterances, and to correlate the results with demographical data of the test persons. Due to the technical problems this goal could not be fully achieved.

However, sufficient data was collected for a first analysis of the spelling recognition performance. Note that only type 1 and 2 spellings were analyzed because the number of elements in both the spelling and the normalized transcription are identical. For type 3 spellings, an alignment is necessary.

The confusion matrix in Table 2 shows that the ratio of correct recognition vs. the highest value for incorrect recognitions was between 20 and 80 for most letters. For *F*, *S*, *M*, *N*, and *AE* this ratio is much lower: 2 for *S*, 5 for *F*, 8 for *M*, and 15 for *N* and *AE.*

Letter pairs	Phoneme Class	Count
T-P, P-T, D-B, D-C, W-B, C-T, P-W, B-W, T-OE, T-A, E-G, B-D, C-OE	plosive	206
F-S, S-F	fricative	182
E-I, E-AE, I-E, I-UE, U-UE, OE-UE, A-AE, R-AE, AE-E	vowel	172
Y--, Y-UE, S-SZ, W--, K-Y, SZ-W, A-*	other	87
N-M, M-N	nasal	62
L-I, AE-L	glide	29
A-H, H-A	aspirated vowel	28
O-U	rounded vowel	26

Table 3: Incorrectly recognized letters by phoneme class

Table 3 contains a phonemical classification of the letter pairs for incorrectly recognized letters that occur 10 and more times in both spelling by letter and by name. Normalized by the number of different letter pairs, *F-S* and *S-F* are incorrectly recognized most often. The spelled letter *F* is pronounced as `/E f/`, and *S* is pronounced as `/E s/` (using the SAM-PA alphabet for German [4]). In German, only the pronunciation of the spelled letters *L*, *N*, *M*, and *R* begin with `/E/`, but in these cases the second component is quite different from `/s/` or `/f/`. Clearly, the reduced signal quality of the telephone line, where the high frequencies are cut off, prevents distinguishing *F* and *S.*

N and *M* are only confused with each other, just like *F* and *S*. Here the reason is that in casual speaking style pronunciation is not very clear, so that the distinction between *N* and *M* becomes blurred. However, the distance to all other letters is sufficient to prevent mistaking a nasal for any other letter.

In German, the spelled letters *B*, *C*, *D*, *G*, *P*, *T*, and *W*, are pronounced as the corresponding plosive or fricative followed by `/e:/`. Hence, recognizing an `/e:/` preceded by something narrows down the choice to these 7 letters. Again, due to the reduced signal quality, and possibly also due to coarticulation in casual speech, the differences between voiced and voiceless or between fricative and plosive disappear, leading to incorrect recognition.

In German spelling, there exist 8 vowels: *A*, *E*, *I*, *O*, *U*, *AE*, *OE*, and *UE*. Phonemically, these vowels can be grouped into two classes: *A*, *AE*, *E*, *I*, *OE*, *UE*, and *O*, *U*. Each of these classes can be seen as a continuum. The region a letter occupies in this continuum cannot be determined exactly, especially in casual speech.

The letter *H* is spelled `/h a:/`. If the aspiration is missing for *H*, or if an *A* is spelled with aspiration, both letters are confused easily.

The pairs *L-I* and *K-Y* are the result of errors in reading the SpeechDat(II) prompt sheet and the subsequent transcription: lower case *L*, upper case *I* (and also the digit *1*) and upper case *K* and *Y* each have a similar graphical shape. Speakers could misread them easily. During the transcription, where the original prompt was displayed on the screen, the transcriber did not notice this mistake because he or she checked the signal with the original prompt before converting it to upper case letters for storage.

Finally, the class *other* contains spellings that can be attributed to incorrect test person input or some non-standard spelling by the speaker. The letter *Y* is spelled as `/Y p s I l O n/`, and this spelling is quite different from that of any other letter. *Y* is a rare letter, and speakers may have used different names or just produced the sound `/Y/` instead of the letter name. The pairs with *SZ* are the result of incorrect data entry (*S* and *Z* separated by a blank instead of *SZ*) or an ambiguity for the German *ß* which has several names: *scharfes s* and *Ess-Zett*, to name the most common expressions. The second form can be confused with the letter sequence *S Z* (spoken as `/E s/ /ts E T/`, which occurs in artificial words or abbreviations.

If spelling by name is used, the confusion of letters because of phonemical similarity is eliminated effectively: *F* and *S* are never confused, and the same is true for *N* and *M*. However, spelling by name introduces new problems: often speakers produce a word beginning with the corresponding letter in an ad hoc fashion. This sometimes leads to wrong words, especially for rare letters and umlauts; also, no word begins with *ß*.

5. CONCLUSION AND OUTLOOK

Due to the technical problems, the data collected was not sufficient to analyze the influence of demographic factors on the recognition performance. However, it is safe to say that humans can recognize spelled letters with an accuracy of approximately 85%. Spelling by name does not improve the overall accuracy, but eliminates mis-recognition due to phonetic similarity of spelled letters.

The experiment will be continued, but with a more robust setup and with only half as many stimuli. Also, both artificial spelling sequences and spellings of real words and names will be used.

The spelling experiment can be found in the WWW at the location *http://www.phonetik.uni-muenchen.de/Spelling.html*

REFERENCES

[1] Draxler, Chr. (1997) WWWTranscribe - A Modular Transcription System based on the WWW, Eurospeech 97, Rhodes

[2] Höge, H., Tropf, H., Winski, R., van den Heuvel, H., Häb-Umbach, R., Choukri, K. (1997) European Speech Databases for Telephone Applications, ICASSP 97, Munich

[3] Schiel, F., Draxler, Ch. & Tillmann, H. G. (1997): The Bavarian Archive for Speech Signals: Resources for the Speech Community; in: Proceedings of the EUROSPEECH 1997, Rhodos

[4] Wells, J. (1998) Standards, Assessment, and Methods: Phonetic Alphabets, http://phon.ucl.ac.uk/home/sampa/home.htm