Natural vs. Synthesized Speech in Spoken Dialog Systems Research – Comparing the Performance of Recognition Results

Tatjana Scheffler[1], *Roland Roller*[1], *Florian Kretzschmar*[2], *Sebastian Möller*[2], *Norbert Reithinger*[1]

[1]DFKI GmbH, Projektbüro Berlin, Alt-Moabit 91c, 10559 Berlin, Germany
[2]Deutsche Telekom Laboratories, TU Berlin, Ernst-Reuter-Platz 7, 10587 Berlin, Germany
Email: [1]firstname.lastname@dfki.de, [2]firstname.lastname@telekom.de
Web: [1]http://www.dfki.de, [2]http://www.qu.tlabs.tu-berlin.de

Abstract

In this paper, we test the effect of using speech synthesis when interacting with a spoken dialog system (SDS). We use a user simulation to connect our speech synthesis to a real, state-of-the-art automatic speech recognition (ASR) component deployed in a working commercial SDS via a standard telephone line. In a series of experiments, we compare human-machine dialogs and their recognition scores with simulated dialogs using synthesis. Our results show that a good text-to-speech synthesis configuration rivals human speech both in recognition scores as well as variability. This makes the speech interface in user simulation quite attractive.

1 Introduction

User simulations for spoken dialog systems can be used for training or evaluation, and there is widespread interest in this approach. However, in most setups, the interaction between the simulated user and the dialog system does not match the intended interface between the completed, deployed system and its eventual human users. Instead of this intended interface of telephone speech, other interfaces such as text or even "intentions" (usually frames or other kinds of extremely task-dependent semantic representations) are used.

There are many reasons for this fact. Most importantly, text or frame/dialog act level output is much easier to produce and interpret than speech output. But using text or intention level interfaces in user simulations carries its own disadvantages, which speak for the use of the speech interface. For one, any telephone-based spoken dialog system can in principle be tested with a user simulation with a speech interface, since one can just access the existing telephone line and use the system's normal ASR. For other interfaces, a special connection (text or intention level) needs to be set up (this may only be possible with special research systems) and extra care needs to be taken to generate simulated ASR errors and other ASR output (such as confidence scores) that the system may use. In case of interfacing at the intention level, the user simulation and system must agree on a pre-specified set of intentions that can be communicated, limiting the variability of the dialog.

On the other hand, there has been doubt that using synthesized speech in user simulation would lead to realistic dialogs. The suspicion was that synthesized speech in combination with ASR is also deterministic, in that the same output is either always correctly recognized or misrecognized in the exact same way. This would not correspond to the effect of real user input, which varies in quality and introduces randomness into the ASR results. To our knowledge, these potential technical disadvantages of the speech interface (combining synthesis with ASR in user simulation for SDS) have not been tested or actually documented before, however.

In SpeechEval [1], we are investigating the option of using a speech interface in user simulations for evaluating dialog systems. To our knowledge the first major evaluation study of a dialog system using speech in the user simulation is [2]. However, in this study, recorded human user utterances were hand-segmented and annotated and the snippets were later re-used in the simulated dialogs. For each planned simulated utterance, a recording from the corpus was chosen as output to the spoken dialog system. For input to the user simulation, text was used instead of speech. In order to achieve the most flexibility (in the output) as well as in order to be able to interact with any given (research or commercial) spoken dialog system in real time, we want to use real speech synthesis and recognition in the user simulation's interface. In this paper, we address specifically the issue of how speech synthesis and speech recognition interact if both are fed immediately into each other (via a telephone line) as in our setup.

The main focus of this paper is how the use of speech synthesis in a user simulation affects the evaluation of the SDS. Therefore, our tests are not lab experiments where speech synthesis is directly connected to an ASR component (maybe even within the same computer). Instead, we use a working end-to-end user simulation to connect our speech synthesis to a real, state-of-the-art ASR component deployed in a working commercial SDS via an ISDN telephone line. This way, realistic transmission degradations can be observed in the tests.

The remainder of this paper is structured as follows. In Section 2, we describe the human user tests which we took as the baseline for our simulated user experiments. Section 3 discusses the setup and results from four sets of experiments we conducted with a simulated user and synthesized speech. A brief discussion and conclusion follows.

2 Human User Experiments

To gather data for comparison with synthetic speech, we conducted an experiment with real users. For this and the following experiments we had access to a separate test line for a commercially deployed German spoken dialog system in the bus information domain. The test line gave us access to the internal logs, including the speech recognition logs, of the system, and was kindly provided by the developing company VoxGen. The SDS uses Nuance version 8.5 for recognition. Nuance ASR is the most-used speech recognition engine for real-life (commercial) German spoken dialog systems, being employed in about 75% of systems[1].

[1]Number among the systems in the VOICE Awards SDS competition in the most recent years 2008 and 2009.

task nr.	from	to	time
1	hotel in Altona	Norderstedt Mitte	leave in one hour
2	Hamburg airport	Feuerbachstr. 2	arrive in two hours
3	Stephansplatz	Rosmarinstr. 5	leave immediately
4	metro stop Baumwall	metro stop Ritterstraße	arrive in one hour

Table 1: Summary of the four tasks, as communicated to human users.

	utterance	recognition result	conf.
SDS:	Okay, sagen Sie mir bitte Ihren Startort.		
	Okay, please state your departure place.		
User:	U-Bahnhof Baumwall.	u bahnhof baumwall	72
SDS:	Okay, nennen Sie mir Ihr Ziel.		
	Okay, state your destination.		
User:	U-Bahnhof Ritterstraße.	u bahnhof ritterstraße	73
SDS:	Gut, und nun noch den Tag und die Uhrzeit.		
	And now please the day and time.		
User:	9. April, Ankunft 15 Uhr.	neunter april ankunft fünfzehn uhr	70
	9. April, arrival 15 o'clock.	*ninth april arrival fifteen o'clock*	
SDS:	Alles klar, . . . // *Alright, . . .*		

Table 2: Excerpt from a human user dialog.

In the user experiment, the subjects first filled out a pre-test-questionnaire asking for demographic data and the subject's attitude towards technology. After that, the users were asked to perform an initial dialog to familiarize themselves with the system without being recorded or having to rate it directly. Afterwards they were asked to perform four different tasks and rate the system after every task on a short questionnaire consisting of 8 items. These dialogs were recorded on the mobile phone for later transcription.

The connections for the tasks were chosen to be not too complicated and error prone, but still varying in complexity. The tasks are summarized in Table 1. We used station as well as street names as places of departure and arrival, and used departure as well as arrival times. We avoided street and station names that were not unique in the cities[2], and tested them beforehand for being not too easily misunderstood by the ASR.

19 human subjects were recorded. Since not all of them managed to complete all four tasks in the alloted time (though all users completed tasks 1 and 2), this procedure led to a total of 60 recorded dialogs. The recordings were subsequently hand transcribed and linked to the corresponding SDS recognition log files to obtain for each user turn the system's recognition result and confidence score. An excerpt from a typical dialog with annotations is given in Table 2. The recognition results for the user tests are summarized in Table 3. Confidence scores were only taken into account for accepted utterances (i.e., utterances that did yield some recognition result, although not necessarily the correct one). The number of user turns in this table (as opposed to the later ones) includes user speech events that the dialog system did not even notice, and for which the recognition was not started (e.g., barge-in attempts when barge-in wasn't allowed).

total number of user turns:	804
accepted utterances:	738
average confidence score:	73.65%
avg. confidence on yes/no-turns:	80.03%
avg. confidence on content turns:	68.75%

Table 3: Recognition results for human user tests.

3 Speech Synthesis Experiments

In the second set of experiments, we carried out the same tasks as in the human user experiment using a simulated user with speech synthesis. The goal was to compare the recognition results and overall effects on the dialogs by switching from real users to synthesized speech. For these tests we used two German speech synthesis engines: a commercial solution by SVOX, and the open source system MARY TTS, developed by DFKI. For both syntheses, we had several different voices and parameter settings available. The dialog system stayed exactly the same as in the user experiment, employing Nuance ASR. The setup for these experiments is shown in Figure 1.

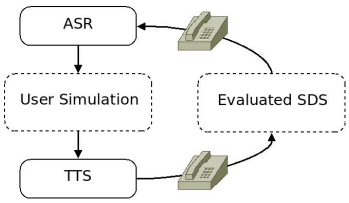

Figure 1: System setup.

3.1 Comparison of Syntheses and Parameters

In a preliminary test, we ran both TTS systems (SVOX and MARY) in different configurations and choosing different voices with a fixed, pre-determined dialog structure in order to test for the best parameter settings overall. The SDS-side recognition results depend heavily on settings such as the volume and speed of the TTS system. Thus, in this test, we defined a fixed walk-through through the dialogs for each of the four tasks (as in Section 2 above). These same four dialogs were then run using a range of settings for

[2] *Feuerbachstraße* from task 2 exists in two cities in the metropolitan area, and users often had to correct the system in its choice of city.

TTS	voices	volume	speed
MARY	bits3 (male) de7 (female) dfki-pavoque-neutral (m.)	40–300	100
SVOX	gl0co0de-DE22 (f.) ag5co0de-DE22 (m.)	50–130	80–100

Table 4: Parameter settings for comparing SVOX and MARY TTS systems.

TTS	voice	vol.[3]	speed	acc.	conf.
MARY	pavoque	100	100	97.9%	**76.4%**
SVOX	female	120	100	94.0%	69.1%
MARY	bits3	110	100	97.1%	66.7%
SVOX	male	100	100	92.9%	65.3%
MARY	de7	110	100	93.8%	64.7%

Table 5: Best configurations of TTS systems according to the comparison experiment.

volume and speed and using each of the available voices in each of the two TTS systems. Table 4 shows the available voices for each system and the range of different settings that were used in this stage of the tests. MARY TTS did not allow for a variation in speed.

In order to compare the various TTS voices and settings we extracted the task completion status (i.e., was a bus connection output by the system?) for the dialog as well as recognition correctness and confidence scores for each simulated user turn. Taking into account only those configurations with the most successful dialogs for each voice, we ranked the TTS systems by overall recognition confidence and acceptance rates (i.e., how many of the turns yielded recognition results). Only the best configurations for each system were retained to be further evaluated in the following experiments. These configurations are shown in Table 5.

3.2 Optimizing the Pronunciation of Named Entities

In a second preliminary test, we hand-optimized the TTS pronunciation for some of the named entities in the four tasks. Some of the street and place names were not in the TTS lexica and could apparently not be segmented correctly by one or both of the TTS systems. For these words we introduced lexicon entries with the correct phonemic representations to see if an increase in recognition confidence scores would result and whether it would lead to more successful or shorter dialogs.

A direct comparison of the recognition rates for only those turns where we hand-tuned the pronunciations (the place names) showed a noticeable improvement for most TTS configurations. As an example, the average recognition scores for the word “Altona” and for each system before and after the optimization are shown in Figure 2.

However, it is interesting to see if hand-tuning the lexicon improves the overall quality of the simulated dialogs. Table 6 shows the effect of the pronunciation optimization on the task success, dialog length and overall average recognition confidence for each TTS configuration for the two tasks (1 and 3) that contained difficult-to-pronounce words. It can be seen that the recognition confidence over all turns increased for all configurations except MARY-pavoque (this TTS did not mispronounce even before the hand-editing). For the two configurations with the lowest success rates (MARY-bits3 and SVOX-male, which each almost entirely failed to ever complete one task because of problems phonemicizing a place name) the success rates were significantly improved by fixing this issue. On the other hand, the average number of user turns for successful calls did not change. This suggests that the pronunciation improvement was not enough to completely avoid confirmation requests by the SDS for the edited place names.

[3]Specified parameters for volume and speed are relative to the pre-set default values for each voice, which equal 100.

3.3 User Simulation with Synthesis

The purpose of the final test was to compare the recognition rates between human user dialogs and simulated user dialogs (with synthesis) in a realistic setting. We set up the simulation to match the recorded user dialogs closely in order to as much as possible abstract away from dialog management or generation module decisions (planning, lexical choice) of the user simulation. A completely free and flexible user simulation may have chosen such different outputs from the real users that any effects of the actual TTS (as opposed to human speech) would be obscured.

The user simulation helps make this kind of comparison possible because a small difference in the recognition result for one input could lead to large differences in the dialog further on if it necessitates additional confirmation or correction turns. Using the user simulation setup, we can recover from such errors and return to the planned dialog protocol automatically. A Wizard-of-Oz setup (typing answers to be synthesized) is not possible with standard dialog systems because the reaction time would be too slow.

In order to replicate the user dialogs closely, we constructed a dialog flow automaton for each task based on the transcriptions of all of the user dialogs using methods from the SpeechEval project [3]. The system prompts were represented as states and the user responses as transition arcs in the automaton. This automaton defines the range of options for our simulation: at each prompt (state), we choose one of the user responses that was given at this point in the dialog by our human subjects. We had to add some unseen states (e.g., additional confirmation requests) in order to make the user simulation more robust. With this setup, a single simulated dialog may not correspond to any particular recorded human dialog. On average, however, the corpus of simulated dialogs covers the same range of inputs and dialog structures as the corpus of human dialogs.

Table 7 compares the resulting simulated dialogs using synthesis with the user dialogs reported on in Section 2, separated by task. Since short “yes”- or “no”-answers are recognized exceptionally well, we report separate confidence scores for confirmations and all other utterances. The results show that the recognition confidence scores for synthesized speech match the scores for human speech very closely, for accepted utterances. A significant difference can however be seen in the success rates and the average length of successful dialogs. In our experiments, even human users had problems solving some of the tasks (2 and 3). This was due to consistently misrecognized words, seen in the high word error rates (WER). The much higher failure rates for the simulated dialogs are largely not due to recognition failures (after all, recognition rates are comparable to humans). But the simulation setup was still not robust enough, since it was based on the dialog flow au-

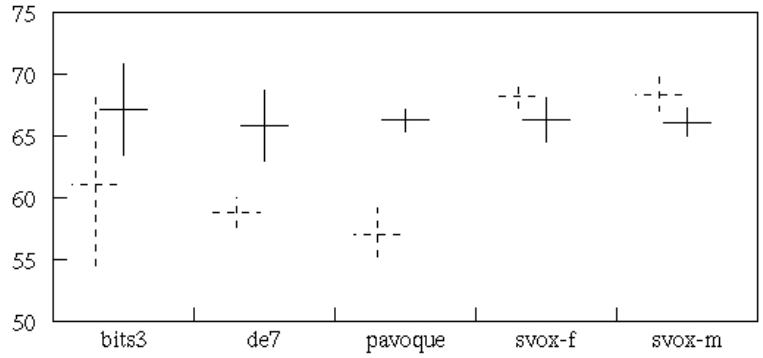

Figure 2: Spread and average recognition scores for "Altona" before (dashed) and after (solid) hand-tuning.

TTS	voice	success		user turns		overall confidence	
MARY	bits3	.13	.63	9.00	9.3	64.43%	66.95%
MARY	de7	.50	.47	9.50	12.63	64.11%	66.31%
MARY	pavoque	.79	.76	8.50	9.13	77.12%	76.43%
SVOX	female	.50	.56	8.75	8.80	65.45%	68.67%
SVOX	male	.38	.55	9.00	8.25	63.11%	66.75%

Table 6: Effect of hand-tuning NE pronunciations. Left columns are before, right columns after tuning.

task	caller	success	user turns	WER	overall conf.	yes/no-conf.	content conf.
1	human users	100%	11.32	14.90%	75.63%	80.63%	70.81%
	MARY pavoque	95.65%	9.61	8.84%	76.70%	82.80%	68.39%
2	human users	75%	15.75	27.96%	72.46%	80.50%	67.34%
	MARY pavoque	38.10%	9.00	26.59%	72.43%	81.37%	64.49%
3	human users	82%	12.27	30.89%	73.60%	80.35%	68.44%
	MARY pavoque	76.19%	9.95	21.99%	73.49%	81.25%	66.27%
4	human users	100%	10.20	25.88%	73.09%	77.07%	69.89%
	MARY pavoque	62.50%	9.50	12.08%	74.91%	81.74%	69.44%

Table 7: Recognition scores and dialog measures for comparable user and simulated dialogs.

tomaton of only up to 20 dialogs. When the simulation hit previously-unseen states, it could not always recover to finish the task (this is why the average dialog length for successful dialogs is *longer* for humans). This problem could be partially solved by making the user simulation more flexible again (not linking it to the previously recorded human corpus as in this experiment). In addition, there were technical problems with the recognition grammar on the user simulation side (esp. for task 2).

4 Conclusions

In this paper we tested the effect of using a speech interface with synthesis in a user simulation. We first evaluated different available TTS systems and their settings. This showed two things. First, there is a large range in the possible recognition confidence scores for different TTSs, with the best ones (in our case, MARY-pavoque) matching human recognition rates. Second and not surprisingly, certain named entities may cause trouble even for good TTS systems. A small hand-edited user lexicon of phonemic representations can help alleviate this problem.

Finally, the direct comparison of user and synthesis calls showed that the recognition confidence scores are in all cases equivalent. This suggests that given a suitably robust user simulation, synthesized speech should not lead to problems in the dialog flow significantly more often than for human callers. We take this as a positive encouragement for our work on user simulation with synthesis.

5 Acknowledgement

This work was part of the IBB project "SpeechEval: Automatic Evaluation of Interactive Speech-Based Services on the Basis of Learned User Models", carried out by DFKI and the Quality and Usability Lab at Technical University Berlin. SpeechEval was funded by the IBB through the ProFIT framework, grant #10140648, and cofinanced by the European Regional Development Fund.

References

[1] S. Möller, R. Schleicher, D. Butenkov, K.-P. Engelbrecht, F. Gödde, T. Scheffler, R. Roller, and N. Reithinger, "Usability engineering for spoken dialogue systems via statistical user models," in *First International Workshop on Spoken Dialogue Systems Technology (IWSDS 2009)*, (Kloster Irsee, Germany), December 2009.

[2] R. López-Cózar, A. de la Torre, J. Segura, and A. Rubio, "Assessment of dialog systems by means of a new simulation technique," *Speech Communication*, vol. 40, pp. 387–407, 2003.

[3] T. Scheffler, R. Roller, and N. Reithinger, "Semi-automatic creation of resources for spoken dialog systems," in *KI 2009: Advances in Artificial Intelligence* (B. Mertsching, M. Hund, and Z. Aziz, eds.), vol. 5803 of *LNAI*, pp. 209–216, Springer, 2009.