ASSESSING THE ACCEPTABILITY OF THE SMARTKOM SPEECH SYNTHESIS VOICES

Antje Schweitzer [1], *Norbert Braunschweiler* [1,2], *Grzegorz Dogil* [1], *Bernd Möbius* [1]

[1] Institute of Natural Language Processing, University of Stuttgart
[2] Rhetorical Systems Ltd, Edinburgh

ABSTRACT

The acceptability of the synthetic voices used by the multimodal SmartKom dialog system was tested in a series of experiments. Early in the project a first set of evaluation tasks was carried out to verify the intelligibility of the diphone voice which serves as the default voice for external open-domain applications. The tests confirmed that the diphone voice produced satisfactory intelligibility. The speech corpus for the unit selection voice recorded by the same speaker is tailored to the typical, more restricted, SmartKom domains. Evaluation tasks focusing on typical SmartKom scenarios demonstrated the superiority of the unit selection voice. In tasks involving open-domain material, however, intelligibility of the unit selection voice appears to be less consistent than that of the diphone voice. In an audio-visual assessment task involving SmartKom specific contexts, the unit selection voice was found to be very well accepted and judged to be satisfactorily intelligible.

1. INTRODUCTION

The task of the speech synthesis group in the SmartKom project [1] was to develop a natural sounding synthetic voice for the avatar, Smartakus, that is judged to be agreeable, intelligible, and friendly by the users of the system.

Two aspects of the SmartKom scenarios facilitate the achievement of this goal. First, since speech output is mainly intended for the interaction of Smartakus with the user, most of the output corresponds to dialog turns generated by the language generation module. As a consequence, most speech output can be generated from linguistic concepts (concept-to-speech synthesis, CTS) produced by the language generation module instead of from raw text (text-to-speech synthesis, TTS). The advantage of CTS over TTS is that it avoids errors that may be introduced by linguistic analysis in TTS mode. Second, the CTS approach narrows down the SmartKom synthesis domain from a theoretically open domain to a restricted domain, which makes unit selection synthesis a promising alternative to diphone synthesis for the SmartKom application.

Multimodality introduces additional requirements for the synthesis module. The visual presence of Smartakus on the screen during speech output requires lip synchronization. Furthermore, Smartakus executes pointing gestures related to objects which are also referred to linguistically. These pointing gestures influence the prosodic structure of the utterance and necessitate temporal alignment of the gestural and linguistic modes. Another momentous requirement was that the graphical design of Smartakus was given before the voice database was recorded. This entailed that the appropriateness of the speaker's voice for Smartakus was an important factor in the speaker selection process.

In developing the synthesis voice for Smartakus, we pursued the following strategy: after the speaker selection process, a diphone voice was developed first using the MBROLA engine [2]. This voice served as a starting point for implementing a unit selection voice by the same speaker tailored to the typical SmartKom domains, and it continues to serve as the default voice for external open-domain applications that require TTS instead of CTS. Both the diphone voice and the unit selection voice were evaluated in the progress of the project.

In this paper we report on a series of experiments that aimed to assess the acceptability of the SmartKom speech synthesis at different points in time during the project. The first set of evaluation tasks was carried out early in the project to verify the intelligibility of the diphone voice. The new unit selection voice used for the more restricted SmartKom scenarios was evaluated shortly before the end of the project. Before presenting the evaluation procedure and results, we first provide some background information on the SmartKom domain and the particular strategies that were implemented to meet the pertinent requirements with an appropriate synthetic voice.

2. SPEECH SYNTHESIS IN SMARTKOM

The SmartKom domains are restricted but not limited: utterances are generated from a number of lexicalized partial syntactic trees [3], but open slots are filled with names, proper nouns, movie titles, etc., from dynamically changing external and internal databases. The vocabulary is therefore unlimited, although it is biased toward domain specific material. The predominance of domain specific material calls for a unit selection approach with a domain specific speech database to ensure optimal speech synthesis quality for frequent phrases. However, since the vocabulary is theoretically unlimited, domain independent material must be taken into account as well. This is especially important because the vocabulary shows typical LNRE (Large Number of Rare Events) characteristics [4]: although each infrequent word on its own is very unlikely to occur, the probability of having an arbitrary infrequent word in an utterance is very high.

Domain specific and domain independent materials pose different requirements for the unit selection strategy. Domain specific phrases may often be found in their entirety in the database. In this case, it may be unnecessary to even consider candidates made up of smaller non-coherent units. Domain independent material, on the other hand, will usually have to be concatenated from much smaller units, such as single segments, demanding a carefully designed database with optimal coverage and a selection algorithm that can handle larger amounts of possible candidates. Therefore, a hybrid approach was implemented combining two existing strategies [5, 6].

2.1. Unit selection

The LNRE characteristics of the SmartKom vocabulary with a limited number of very frequent domain specific words and a large number of very infrequent words originating from dynamic databases suggested a hybrid unit selection strategy that integrates two well-known methods, bottom-up acoustic clustering (AC) [7] and top-down phonological structure matching (PSM) [8].

The AC algorithm achieves a reduction of unit candidate sets by clustering all units in the database according to their linguistic properties in such a way that the acoustic similarity of units within the same cluster is maximized. During synthesis, the linguistic context determines the appropriate cluster. Unit candidate sets are typically very large for frequent units, and the number of possible sequences of candidates grows dramatically with the number of candidates. For performance reasons, the candidate sets must therefore be reduced, although at the risk of excluding originally adjacent candidates. The AC procedure enhances the efficiency of a bottom-up unit selection approach in which, starting from the segmental level, the selection of complete syllables, words or phrases arises indirectly as a consequence of lower concatenation costs for adjacent segments.

In the PSM algorithm, candidates are searched top-down on different levels of the linguistic representation of the target utterance. If no candidates are found on one level, the search continues on the next lower level. If appropriate candidates are found, lower levels are ignored for the part of the utterance that is covered by the candidates. This approach is primarily designed for limited domains, where it benefits from the fact that most longer units are represented in the database. The advantage of such a top-down approach is that it favors the selection of these longer units in a straightforward way. If candidates are found on levels higher than the segment level, this strategy can be faster than the bottom-up approaches because there are longer and therefore fewer unit candidates. Still, particularly on the segment level, candidate sets may be very large.

In the hybrid unit selection method applied in the SmartKom system, the PSM strategy ensures high-quality synthesis for frequent material by directly selecting entire words or phrases from the database. If no matching candidates are found above the segment level, which will typically be the case for domain independent material, the AC approach serves to reduce the amount of candidate units. Note that the SmartKom implementation of the PSM algorithm differs from the original implementation [8] in several aspects; the details can be found in [5, 6].

2.2. Corpus design

The requirements for the contents of the database are again different for domain specific vs. domain independent material. For the limited amount of domain specific material, it is conceivable to include typical words in several different contexts [9] or even to repeat identical contexts. In contrast, for the open-domain part a good coverage of the database in terms of diphones in different contexts is essential, as emphasized by [10, 4].

We applied a greedy algorithm to select from a German newspaper corpus of 170 000 sentences a set of utterances that maximized coverage of units [10]. We built a feature vector for each segment including its phonemic identity, syllabic stress, word class, prosodic and positional properties. Additionally, the diphone sequence for each sentence was determined. Sentences were then selected successively by the greedy algorithm according to the number of both new vectors and new diphone types that they covered. For German diphone types that did not occur at all, sentences were constructed that would contain them. These sentences were added to the corpus, and the selection process was repeated. This ensured that at least a full diphone coverage was obtained, and at the same time the number of phoneme/context vector types was increased.

We added 2643 SmartKom specific words and sentences to the domain independent corpus. They included excerpts from demo dialogs, but also domain typical slot fillers such as proper names and place names, numbers, weekdays, etc. Movie titles, many of them in English, constituted the largest group of domain specific material, partly to make up for the omission of English phones in the systematic design of the text material. The speech database was recorded by the same professional speaker as the diphone voice and amounts to about 160 minutes of speech (Table 1).

length	sentences	words	syllables	segments	speakers
160 min.	2 600	17 400	33 800	94 300	1

Table 1. Size and structure of the SmartKom unit selection speech database.

	first evaluation		second evaluation			
			pilot study		full experiment	
	material	voice	material	voice	material	voice
dictation	SUS	natural			SUS	diphones
		diphones				unitsel
	SmartKom	diphones				
listening comprehension					open domain	diphones
						unitsel
subjective impression	SmartKom	diphones	SmartKom	diphones	SmartKom	unitsel
				unitsel		

Table 2. Overview of the tasks performed by participants in the evaluation procedures. The general type of task is indicated in the leftmost column. The table lists the type of text material, viz. normal text (open domain), semantically unpredictable sentences (SUS), or SmartKom specific material (SmartKom), and the voices used to generate the stimuli, viz. diphone voice or unit selection voice.

3. EVALUATION: GENERAL PROCEDURE

Due to the complexity of multimodal systems, it is difficult to evaluate single components because they are not designed to perform in a stand-alone mode, isolated from other system components that they interact with. Also, the performance of the system as a whole, not the performance of its modules, is decisive for user acceptance and usability. As a consequence, the SmartKom system has been subjected to an end-to-end system evaluation [1].

However, an additional evaluation of the speech synthesis component is necessary to give more detailed, possibly diagnostic, insights into potential synthesis specific problems. This can be difficult since the boundaries between system components are often not clear-cut from a functional point of view. In SmartKom, language generation and synthesis are strongly linked. Without language generation, simulating concept input for CTS synthesis is tedious. But if concept input is generated automatically for synthesis evaluation purposes, the language generation component is implicitly evaluated together with the synthesis module. A second problem is that the appropriateness of the synthesis voices for Smartakus cannot be evaluated without the animation component.

To detect possible synthesis specific problems, we carried out evaluations of the synthesis module, detached as far as possible from the SmartKom system, at two times. The first evaluation took place early in the project and served to verify that the diphone synthesis voice produced satisfactory intelligibility; the second evaluation was carried out in the last project phase to assess the quality of the new unit selection voice, particularly in comparison to the diphone voice. Table 2 shows an overview of the tasks performed in the evaluation procedures.

4. EVALUATION: DIPHONE VOICE

The first evaluation involved a total of 58 participants, which can be classified in two groups. The first group comprised 39 students of the University of Ulm. These subjects are referred to as “naive” because they reported to have had no prior experience with speech synthesis or language processing. The second group consisted of employees of DaimlerChrysler at Ulm, who were experienced with regard to speech technology. All participants completed three dictation tasks: one with SmartKom specific utterances rendered by the diphone voice, one with semantically unpredictable sentences (SUS) [11] recorded by a speaker, and one using SUS stimuli synthesized by the diphone voice.

4.1. SmartKom dictation task

The SmartKom specific dictation task was intended to verify that the intelligibility of the diphone voice was satisfactory for the use in SmartKom. The participants transcribed nine system turns in a continuous dialog between the system and a user. 93% of these system turns were transcribed without any errors, 4% involved obvious typing errors, and in 2% of

template	constituent	lexical slots
S V O	subject	determiner (sg.) + noun (sg.)
	verb	transitive verb (3rd person sg.)
	object	plural noun
S V PP	subject	determiner (sg.) + noun (sg.)
	verb	intransitive verb (3rd person sg.)
	adjunct PP	preposition + determiner (acc. sg.) + noun (acc. sg.)
PP V S O	adjunct PP	preposition + determiner (dat. sg) + noun (dat. sg.)
	verb	transitive verb (3rd person sg.)
	subject	determiner (nom. sg.) + noun (nom. sg.)
	object	determiner (acc. sg.) + noun (acc. sg.)
V S O!	verb	transitive verb (imperative pl.)
	subject	"Sie"
	object	determiner (acc. sg.) + noun (acc. sg.)
V S O?	verb	transitive verb (3rd sg.)
	subject	determiner (nom. sg) + noun (nom. sg)
	object	determiner (pl.) + noun (pl.)

Table 3. Overview of syntactic templates used for the generation of SUS stimuli. The table shows the lexical slots in the templates corresponding to the constituents in each of the templates. The symbols '!' and '?' indicate imperative and interrogative sentence mode, respectively. Although not explicitly stated here, noun phrases were also congruent in gender, and the complements of transitive verbs and prepositions were in the appropriate case.

the transcriptions there were errors which can probably be attributed to memory problems rather than to intelligibility. These figures show that the diphone voice offers excellent intelligibility for normal speech material.

4.2. SUS dictation tasks

The SUS dictation tasks are perceptually more demanding because the linguistic context does not provide any cues in cases of locally insufficient intelligibility. The tasks thus aimed at testing the intelligibility of the diphone voice under more challenging conditions. The sentences were generated automatically using five different templates, which are listed in Table 3. The material to fill the lexical slots in the templates came from lists of words selected from CELEX [12] according to their morphological and syntactical properties. The lists were randomized before generating the SUS stimuli. Thus, all lexical items were used at least once, but in varying combinations.

The SUS task using natural stimuli immediately preceded the task with the diphone stimuli. It served to estimate the upper bound of scores in such a task. The subjects transcribed 15 stimuli in each of the two tasks. For the natural stimuli, the sentence error rate was 4.9%. Of these, 0.6% were obvious typing errors. The error rate for the synthesized stimuli was 33.9%. Again, 0.6% were typing errors.

The error analysis for the diphone stimuli showed three relatively frequent error types. One concerned the confusion of short and long vowels. This can probably be attributed to the duration model used for determining segmental durations, which had been trained on a speech corpus from a different speaker. We replaced this model with a speaker specific model trained on the unit selection voice data later in the project. Another problem was that sometimes the subjects did not correctly recognize word boundaries. We expect that in these cases listeners should also benefit from the improved duration model. The other two types of errors concerned voiced plosives preceding vowels in word onsets, and voiced and voiceless plosives preceding /R/ in the same position. We claim that the latter is a typical problem in diphone synthesis: the two /R/-diphones concatenated in these cases are two different positional variants of /R/, viz. a postconsonantal variant, and an intervocalic variant.

4.3. Subjective assessment

After performing the dictation tasks, participants were asked for their subjective impression of the diphone voice. They rated the voice on a five-point scale ranging from -2 to +2 for each of the two questions *"How did you like the voice?"* (-2 and +2 corresponding to "not at all" and "very much", respectively), and *"Did you find the voice easy or hard to understand?"* (-2 and +2 corresponding to "hard" and "easy", respectively). Subjects also answered "yes" or "no" to the question *"Would you accept the voice in an information system?"*.

The results strongly indicate that non-naive subjects generally rated the voice better than naive subjects. The mean scores for the first two questions broken down by experience with speech technology were +0.53 and +1.37 for non-naive participants, and -0.21 and +0.67 for naive participants, respectively. Of the non-naive subjects, 95% said they would accept the voice in an information system, while only 72% of the naive subjects expressed the same opinion. The first evaluation thus confirmed that the diphone voice yielded satisfactory results.

5. EVALUATION: UNIT SELECTION VOICE

The second evaluation focused on the unit selection voice. Here the diphone voice served as a baseline for the dictation and listening comprehension tasks. The actual evaluation was preceded by a pilot study on the acceptability of the unit selection voice versus the diphone voice specifically for typical SmartKom utterances.

5.1. Pilot study

The subjects in this pilot study were students from Stuttgart and their parents. The younger student group and the older parent group each consisted of 25 participants. Subjects listened to 25 SmartKom specific dialog turns in randomized order, both rendered in the unit selection voice and in the diphone voice. Afterwards, they were asked to answer the questions *"How do you judge the intelligibility of the synthesis voice?"* and *How do you judge the suitability of this voice for an information system?"* on a five-point scale ranging from -2 ("very bad") to +2 ("very good").

There was a similar effect observable between the younger and the older group as in the first evaluation between the non-naive and the naive group. The younger group was more tolerant to diphone synthesis regarding intelligibility: the mean scores for the diphone voice were +0.83 for the younger group and +0.51 for the older one. The unit selection voice was rated significantly better by both groups; the mean score was +1.76 in both cases. The results for the question regarding the suitability of the voices in an information system show that the unit selection voice is strongly preferred. Mean scores were clearly below zero for the diphone voice (-1.21 and -1.33 for the younger and the older group, respectively), and clearly above zero for the unit selection voice (+1.79 and +1.23 for the younger and the older group, respectively).

5.2. SUS transcription task

In the following evaluation, 77 subjects participated, none of which had taken part in the earlier evaluations. Three tasks were completed in this evaluation. Participants first transcribed SUS stimuli. The stimuli were taken from the first evaluation, but they were synthesized using both the diphone and the unit selection voices.

The results are comparable to the earlier results: the sentence error rate was 27% including typing errors for the diphone voice (earlier: 33%). This shows that the diphone voice has gained in intelligibility compared to the first evaluation. For the unit selection voice, however, the error rate was 71%. This is due to the fact that the SUS stimuli contained only open-domain material. The unit selection voice was designed for a restricted domain with prevailing SmartKom specific material. In this respect, completely open domains are a worst-case scenario, in which the synthesis quality must be expected to be inferior to that of SmartKom specific material. Additionally, at the time of conducting the evaluation, the speech database was still in the process of being manually corrected. Informal results obtained at the end of the project, i.e. two months after the formal evaluation and after extensive manual correction of prosodic and segmental corpus annotations, indicate that the subjective synthesis quality especially for open-domain material has improved since the completion of the evaluation.

5.3. Subjective assessment

After completing the SUS dictation task, participants were presented three video clips showing the SmartKom display during a user's interaction with Smartakus. The user's voice had been recorded by a speaker. The system's voice in the video clips was the unit selection voice, synchronized with Smartakus's lip movements and gestures. After this task, subjects were asked to answer three questions by adjusting a sliding bar between two extremes. The three questions were *"How intelligible did you find the voice?"* with possible answers ranging from "not intelligible" to "good", *"How natural did you find the voice?"* with answers between "not natural at all" and "completely natural", and *"How did you like the voice?"* with answers between "not at all" and "very well".

The results for the three answers were 71% for intelligibility, 52% for naturalness, and 63% for pleasantness. These figures show that in the SmartKom specific contexts, the unit selection voice is very well accepted and judged to be satisfactorily intelligible. This confirms the results obtained in the pilot study for audio-only stimuli.

5.4. Comprehension task

In the listening comprehension test, the subjects listened to four short paragraphs of open-domain texts. After each paragraph, they were asked three questions concerning information given in the text. Two texts were rendered using the diphone voice, two using the unit selection voice.

The results were again better for the diphone voice, with 93% of the answers correct, while 83% were correct for the unit selection voice. In this context, both voices were rated lower than in the SmartKom specific task. The scores were 53%, 34%, and 42% for the diphone voice, and 23%, 22%, and 26% for the unit selection voice, respectively. Again, we expect much better results after the manual correction of the speech database.

6. CONCLUSION

To summarize, the superiority of the unit selection voice is evident for the SmartKom domain. This was confirmed by the pilot study and the SmartKom specific part of the second evaluation. The quality of the diphone voice has improved between the first and the second evaluation. We attribute this effect mainly to the new duration model obtained from the unit selection data of our speaker. The ongoing manual correction of the unit selection database is evidently effective. Subjectively, the synthesis quality has improved since the completion of the second evaluation. However, this will have to be confirmed in more formal tests.

Future work will focus on the extension of our unit selection approach from the restricted SmartKom domain to open domains in general. The experience gained in working with the SmartKom unit selection voice suggests that accuracy of the database annotation is crucial for optimal synthesis quality. Also, the strategy to deal with large numbers of unit candidates as they often occur in open-domain sentences without excluding potentially good candidates will need some more attention in the future.

7. ACKNOWLEDGMENTS

The contributions of Martin Ernst (DaimlerChrysler, Ulm) and Gerhard Kremer, Wojciech Przystas, Kati Schweitzer, and Mateusz Wiacek (IMS) to the evaluation sessions are gratefully acknowledged. This work was funded by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the SmartKom project under Grant 01IL905K7. The responsibility for the content lies with the authors.

8. REFERENCES

[1] Wolfgang Wahlster, Ed., *SmartKom—Foundations of Multimodal Dialogue Systems*, Springer, 2004, to appear.

[2] Thierry Dutoit, Vincent Pagel, N. Pierret, F. Bataille, and O. van der Vrecken, "The MBROLA project: Towards a set of high quality speech synthesizers free for use for non commercial purposes," in *Proc. International Conference on Spoken Language Processing (Philadelphia)*, 1996, vol. 3, pp. 1393–1396.

[3] Tilman Becker, "Fully lexicalized head-driven syntactic generation," in *Proc. Ninth International Workshop on Natural Language Generation*, Niagara-on-the-Lake, Ontario, Canada, 1998, pp. 208–217.

[4] Bernd Möbius, "Rare events and closed domains: Two delicate concepts in speech synthesis," *International Journal of Speech Technology*, vol. 6, no. 1, pp. 57–71, 2003.

[5] Antje Schweitzer, Norbert Braunschweiler, Tanja Klankert, Bernd Möbius, and Bettina Säuberlich, "Restricted unlimited domain synthesis," in *Proc. European Conference on Speech Communication and Technology (Geneva)*, 2003, pp. 1321–1324.

[6] Antje Schweitzer, Norbert Braunschweiler, Grzegorz Dogil, Tanja Klankert, Bernd Möbius, Gregor Möhler, Edmilson Morais, Bettina Säuberlich, and Matthias Thomae, "Multimodal speech synthesis," in *SmartKom—Foundations of Multimodal Dialogue Systems*, Wolfgang Wahlster, Ed. Springer, 2004, to appear.

[7] Alan W. Black and Paul Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *Proc. European Conference on Speech Communication and Technology (Rhodos, Greece)*, 1997, vol. 2, pp. 601–604.

[8] Paul Taylor and Alan W. Black, "Speech synthesis by phonological structure matching," in *Proc. European Conference on Speech Communication and Technology (Budapest, Hungary)*, 1999, vol. 2, pp. 623–626.

[9] Karlheinz Stöber, Petra Wagner, Jörg Helbig, Stefanie Köster, David Stall, Matthias Thomae, Jens Blauert, Wolfgang Hess, Rüdiger Hoffmann, and Helmut Mangold, "Speech synthesis using multilevel selection and concatenation of units from large speech corpora," in *Verbmobil: Foundations of Speech-to-Speech Translation*, Wolfgang Wahlster, Ed., pp. 519–534. Springer-Verlag, 2000.

[10] Jan P. H. van Santen and Adam L. Buchsbaum, "Methods for optimal text selection," in *Proc. European Conference on Speech Communication and Technology (Rhodos, Greece)*, 1997, vol. 2, pp. 553–556.

[11] Christian Benoît, Martine Grice, and Valerie Hazan, "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences," *Speech Communication*, vol. 18, pp. 381–392, 1996.

[12] Harald Baayen, Richard Piepenbrock, and Léon Gulikers, "The CELEX lexical database—Release 2," CD-ROM, 1995, Centre for Lexical Information, Max Planck Institute for Psycholinguistics, Nijmegen; Linguistic Data Consortium, University of Pennsylvania.