Multimodal speech synthesis

Antje Schweitzer, Norbert Braunschweiler, Grzegorz Dogil, Tanja Klankert, Bernd Möbius, Gregor Möhler, Edmilson Morais, Bettina Säuberlich, and Matthias Thomae

Institute of Natural Language Processing, University of Stuttgart

The main goal of the speech synthesis group in SmartKom 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 SmartKom system.

Two aspects of the SmartKom scenario 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 (see Chapter **??**). Therefore, most speech output can be generated from linguistic concepts produced by the language generation module ("concept-to-speech synthesis", CTS) 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 that are 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 could be included as 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. This voice served both as a starting point for implementing a unit selection voice by the same speaker tailored to the typical SmartKom domains, and as the default voice for external open domain applications that require TTS instead of CTS. The diphone voice and the unit selection voice were both evaluated in the progress of the project.

This chapter is organized as follows. We focus on the prosody generation in CTS mode in the subsequent section. The speaker selection process is described in Sec-

tion 2. Section 3 concentrates on the unit selection voice. Lip synchronization and gesture-speech alignment are discussed in Section 4. Finally, the two evaluation procedures are described in Section 5.

1 Concept-to-speech synthesis

The motivation for CTS synthesis is the view that the linguistic content of an utterance determines its phonological structure and prosodic properties. It has been shown that prosodic structure can reflect aspects of syntactic structure, information structure, and discourse structure [12, 30, 10, 16, 1, 9, 22]. The challenge in TTS is that text represents only a very reduced version of the full linguistic content of an utterance. It not only lacks marking of higher-level linguistic structure, but may also be ambiguous with respect to syllabic and segmental structure due to abbreviations and homographs. All these properties have to be inferred from the text in TTS. The idea of CTS is to use the full linguistic structure of an utterance, i.e. the original "concept", instead of its raw textual representation. This structure is available in dialog systems which generate utterances dynamically. In SmartKom, it is available with some exceptions: many utterances contain material retrieved from external databases, such as movie titles or geographical names. Although the overall structure of such utterances is known, the internal structure of the retrieved items is unknown. They may contain abbreviations, material in unknown languages, or, particularly in the case of movie titles, may even have their own internal linguistic structure.

The main advantage of CTS in SmartKom is therefore the availability of higher-level linguistic structure, which influences the prosodic structure of an utterance. Cinque [10] gives a detailed account of how syntactic structure determines the default location of sentence stress. We have implemented an algorithm motivated by Cinque's findings. The prediction of prosodic structure including pitch accent and boundary types from linguistic structure is described in more detail in [28]. Here we only give a brief description of the concept structure, the prediction of phrasing, and the implementation of Cinque's account for accent placement.

1.1 Concept input

Concepts in SmartKom contain information on three linguistic levels. The highest level of annotation used for prosody prediction is the **sentence level**. Sentence mode (declarative, imperative, yes/no-question, or wh-question) is annotated on this level. This kind of information is mainly required for the prediction of boundary tones.

The next lower level is **syntactic structure**. Syntactic trees in SmartKom are binary branching, and they may include traces resulting from movement of syntactic constituents. They are generated from smaller tree segments within the tree-adjoining grammar framework [4]. Semantic and pragmatic information is integrated into the syntactic structure as follows. For each node of the syntactic tree, its argument status (subject, direct or indirect object, prepositional object, sentential object, or adjunct) and its information content (new vs. given) can be specified. Deixis is also specified

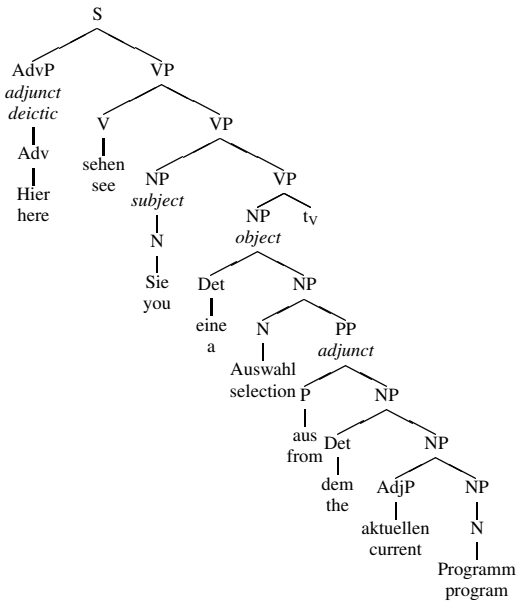


Fig. 1. Integration of additional information into the syntactic and lexical structure of the sentence *Hier sehen Sie eine Auswahl aus dem aktuellen Programm*. In this example, deixis (*deictic*) and argument status (*adjunct*, *subject*, *object*) are added. These values are indicated in italics.

on the syntactic level. Deictic elements occur when Smartakus executes pointing gestures referring to objects on the screen.

The lowest level of annotation is the **lexical level**. On this level, material that originates from database queries is inserted. The domain and language of this material are annotated if available. An example of the syntactic and lexical levels of a concept structure is given in Figure 1.

1.2 Prediction of prosodic phrases

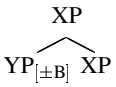
The first step in prosody generation is the prediction of prosodic phrase boundaries. There are two levels of phrases: intonation phrases are terminated by major breaks ("big breaks", BB) and can be divided into several intermediate phrases, which in turn are separated by minor breaks (B).

Syntactic structure has been shown to be useful in the prediction of prosodic phrasing [29]. Particularly the insertion of prosodic phrase breaks between topicalized constituents in the *Vorfeld* (i.e., constituents preceding the finite verb in verb-second sentences) and the rest of the sentence has proved to be a common phenomenon in natural speech, if the material in the Vorfeld is long enough [29]. The Vorfeld corresponds to a syntactic constituent, a maximal projection, that is in the specifier position of another maximal projection (depending on the syntactic theory

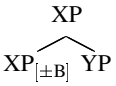
a verbal, inflectional or complementizer projection). Prosodic breaks can also occur between constituents in the *Mittelfeld*. Another observation is that breaks are less likely between heads and complements than between heads and adjunct constituents. In any case, the longer the constituents are, the more likely the breaks are inserted. Usually, the inserted breaks are minor breaks; occasionally, even major breaks occur.

These observations motivate the two rules in (1) and (2), which insert optional minor breaks. The [±B] feature indicates that a break can be inserted at the end of the respective constituent. The rules each have two variants (a) and (b), which are mirror images of each other.

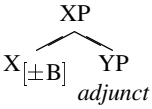
(1) a.



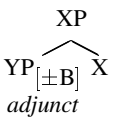
b.



(2) a.



b.



The first rule states that maximal categories (the YPs in (1)) that are daughters of other maximal categories (the dominating XPs in (1)) can be separated from their sister node by a minor break. S constituents are also treated as maximal projections. Since we do not distinguish X-bars from XPs, rule (1) applies to any maximal projection that is not the sister of a head. Examples of the application of (1) are the insertion of boundaries between topicalized constituents and the VP as well as between adjacent constituents within the VP. The second rule allows breaks to be inserted between the head of a phrase (the X in (2)) and its sister node, but only if the sister node is an adjunct. Thus, phrase boundaries between a head and its argument are excluded.

Deictic elements often trigger additional minor phrase breaks. A pilot study on material from 26 speakers showed that deictic expressions, i.e. expressions that were accompanied by pointing gestures, were usually marked by a phrase break or an emphatic pitch accent or both. This effect is modeled by inserting mandatory minor breaks preceding and following deictic expressions.

The result of the phrase break insertion for the sentence in Figure 1 is shown in (3). Mandatory phrase breaks are (trivially) at the end of the utterance, and after the deictic AdvP *hier*, indicated by the [+BB] and [+B] features, respectively. Optional phrase breaks are inserted after the NP *Sie* according to rule (1-a), and after the noun *Auswahl* according to rule (2-a). These optional breaks are marked by the feature [±B] in (3).

(3) Hier [+B] sehen Sie [±B] eine Auswahl [±B]
aus dem aktuellen Programm [+BB]

In a second step, a **harmonization algorithm** selects candidates from the set of possible combinations of prosodic phrases. Candidates whose mean phrase length lies in a given optimal range and which show an even phrase length distribution are favored over other candidates. Thus, the observation that the insertion of breaks depends on the length of the resulting phrases is accounted for, and sequences of phrases that are unbalanced in terms of number of syllables per phrase are avoided if possible. The optimal range for the mean phrase length was found to be more than 4 to less than 11 syllables.

For the example in Figure 1, the optimal candidate is shown in (4-a). The other candidates are given in (4-b) through (4-d). Syllable number per phrase, mean phrase length and variance are indicated in italics. (4-b) is discarded because its mean phrase length is not in the optimal range. From the remaining three candidates, (4-a) is chosen because it has the smallest variance.

(4) a. Hier [+B] sehen Sie eine Auswahl [+B] aus dem aktuellen Programm [+BB]
syllables: 1, 7, 8; mean: 5.33; variance: 9.55

b. Hier [+B] sehen Sie [+B] eine Auswahl [+B] aus dem aktuellen Programm [+BB]
syllables: 1, 3, 4, 8; mean: 4; variance: 6.5

c. Hier [+B] sehen Sie [+B] eine Auswahl aus dem aktuellen Programm [+BB]
syllables: 1, 3, 12; mean 5.33; variance: 22.89

d. Hier [+B] sehen Sie eine Auswahl aus dem aktuellen Programm [+BB]
syllables: 1, 15; mean 8; variance: 49

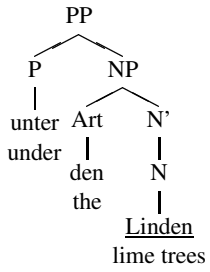
Finally, boundary tones are assigned to each predicted major phrase boundary. For sentence-internal phrase boundaries, a rising boundary tone is assigned to indicate continuation. In all other cases, the boundary tone depends on the sentence mode.

1.3 Accent prediction

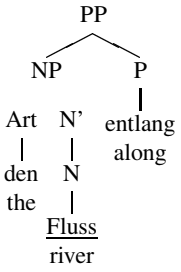
The default location of sentence stress is determined by the syntactic structure according to [10]. We have adapted this procedure to predict the default accent location for each prosodic phrase. Additionally, semantic factors can cause deaccentuation. Pitch accent types depend on the information content of the accented word, on its position in the phrase, and on sentence mode.

According to [10], the default accent is on the syntactically most deeply embedded element, as illustrated by the prepositional phrases in (5) (from [10]). The underlined words are the most deeply embedded elements, and they are accented by default.

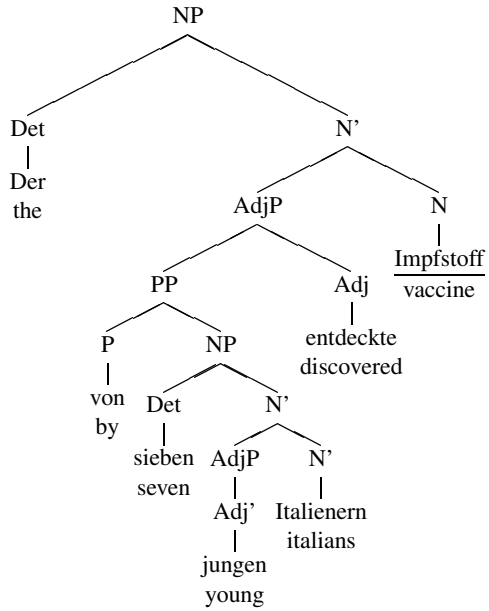
(5) a.



b.



(6)



However, depth of embedding on the non-recursive side is irrelevant, as shown in (6) (from [10]): in neutral accentuation, the pitch accent falls not on the overall most deeply embedded element, *Italienern*, but on *Impfstoff*. This is because NPs are right-recursive. Depth of embedding according to [10] is only counted on a path along the X-bar axis (e.g., connecting XP and X', X' and X') and on the recursive side of each projection XP (e.g., connecting X' to a YP embedded on the left side, if XP is a left-recursive category; or connecting X' to a YP embedded on the right side, if XP is a right-recursive category). The main path of embedding is the path that reaches the top node. The overall most prominent element is the most deeply embedded element on the main path of embedding. In constituents on the non-recursive side, depth of embedding determines the locally most prominent element in the constituent, but its depth of embedding is irrelevant for the location of the main stress.

This procedure had to be modified for two reasons. First, in the syntactic structures used in SmartKom, there are no X-bars. Thus, the main path is along an axis connecting XPs with embedded XPs, or connecting an XP with a maximal projection YP on the recursive side of the XP, if YP is a sister to the head X of XP. Second, large syntactic trees will usually be split into smaller units by the phrase prediction algorithm. In the phrases that do not contain the globally most prominent element according to the definition above, we still need to assign an accent to the locally most prominent element. The resulting procedure works as follows. In each phrase, the element with the smallest number of branches on the non-recursive side is accented. If there are several elements with the same number of branches, the last one is accented. Depending on the information structure of an utterance, accentuation can deviate from the default accentuation: words are deaccented if they are marked as "given" in the respective context, and narrow focus moves the accent from the default location to the focused constituent.

For each accented element, its accent category depends on its position in the phrase, its information content, and the sentence mode. We use a subset of the pitch accent inventory of the German ToBI labeling system as described in [21], viz. L*H as a rising accent, H*L as a falling accent, and L*HL as an emphatic accent. For the diphone voice, the type of accent determines the template used for modeling the fundamental frequency contour [24]. For the unit selection voice, it restricts the candidate set to candidates realized with a similar accent (see Section 3).

1.4 An example

The complete prosody prediction algorithm is illustrated by the example in (7) and (8). An optional phrase break is inserted between the topicalized object *Das Dokument* and the finite verb *wurde*. The harmonization algorithm selects (8-a) because (8-b)'s mean phrase length exceeds the upper limit of 11 syllables per phrase and is therefore not considered in the selection step. Otherwise, it would have been preferred over (8-a) because its variance is smaller.

(7) Das Dokument wurde an Nils Nager verschickt.
The document was to Nils Nager sent
The document was sent to Nils Nager.

(8) a. Das Dokument [+B] wurde an Nils Nager verschickt [+BB]
syllables: 4, 8; mean: 6.00; variance: 4.00
b. Das Dokument wurde an Nils Nager verschickt [+BB]
syllables: 12; mean: 12.00; variance: 0.00

The syntactic structure of (8) is shown in Figure 2. The main path of embedding is indicated by the nodes in bold face. For the first phrase, the default accent is assigned to the noun *Dokument* because the path from the top of the tree to the noun contains only one branch (connecting S to the NP on its left) that is neither on the recursive side nor on the X-bar axis. The branches connecting NP to NP and NP to N are on the recursive side because NPs are right-branching. The path to the determiner *Das*

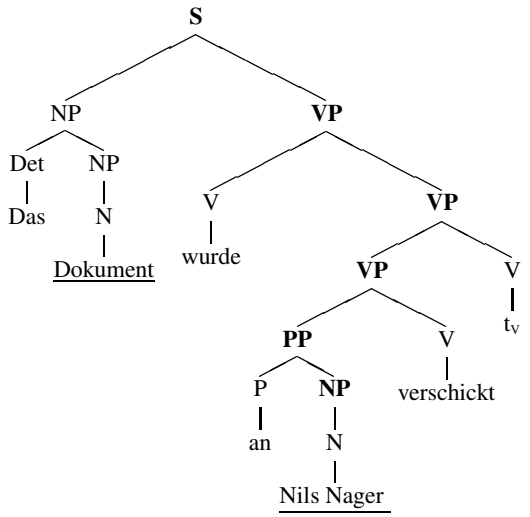


Fig. 2. Syntactic structure for the example "Das Dokument wurde an Nils Nager verschickt" (*The document was sent to Nils Nager*). The main path of embedding is marked by the nodes in bold face. The default accents for the two phrases are assigned to the underlined words.

contains two branches that are on the non-recursive side: the branches connecting S to NP and NP to Det, respectively. In the second phrase, the name *Nils Nager* is on a path exclusively along the X-bar axis or along branches on the recursive side. It is therefore accented.

Since the sentence is a declarative sentence, it is terminated by a falling boundary tone. The accented element in the second phrase is assigned a falling accent for the same reason. The accent in the first phrase is predicted to be rising because the sentence continues across the intermediate phrase boundary between the two phrases. Thus, the prosodic structure for (7) is as shown in (9).

(9) Das Dokument wurde an Nils Nager verschickt.
L*H - H*L L%

2 Speaker selection

Several constraints have to be met in the speaker selection process. On the one hand, users' expectations include not only intelligibility but also more subjective properties such as agreeableness, pleasantness, and naturalness. Adequacy of the voice for the target application may be even more important than subjective pleasantness. For instance, Smartakus is a small blue-colored cartoon-like character reminiscent of the letter "i". This visual appearance did not seem to go well with one particularly deep candidate voice, which was rated high by listeners only when presented independent of Smartakus.

There are additional, more technical and practical, requirements, such as the experience of the speaker, which can decisively reduce the time needed for the recordings, but also foreign language skills, which are required for some non-native diphones, as well as the speaker's availability over a longer period of time.

The subjectively perceived properties of a diphone voice are currently not predictable from the speaker's natural voice. The prediction is less difficult for unit selection voices because they preserve the characteristics of the original voice much better by reducing the number of concatenation points and the amount of signal processing. However, the number of concatenation points may be similar to that in diphone synthesis, in which case the subjective voice quality is almost as hard to predict as in the diphone case. To ensure that the selected speaker's voice is suitable for diphone synthesis in the sense that the resulting diphone voice is still judged to be agreeable, we built a test diphone voice for each speaker. To this end, a small diphone set was recorded that covered the diphones required for synthesizing three short sentences. The speaker with the best test voice was selected.

Since recording and building a diphone database is very time-consuming even for the rather small set of diphones needed for the test voices, we split the speaker selection process in two phases. In the first phase, we asked speakers to record some SmartKom specific material. This material included a short dialog typical of a SmartKom domain, a list of (nonsense) diphone carrier words, and three short excerpts from movie reviews in German, English, and French. Some speakers sent in demo tapes, and some were recorded in an anechoic recording booth at our lab. Altogether, we collected demo material from 40 speakers, 29 female and 11 male. For each voice, some representative sentences were selected and rated for their subjective qualities in an informal evaluation procedure. Most participants in this rating procedure were colleagues from our institute.

In the second phase, the ten best speakers from the first phase, 4 male and 6 female, were invited to our lab to record the diphone set required for our three test sentences. The diphones were manually labeled and afterwards processed by the MBROLA[1] group at Mons. We carried out a formal evaluation with 57 participants; 20 participants were experienced and 37 "naive" with respect to speech technology. The three target sentences were synthesized for each speaker using different signal processing methods (MBROLA [14], PSOLA [26] and Waveform Interpolation [25]) and different prosody variants (the speaker's original prosody vs. prosody as predicted by our TTS system, with the pitch range adapted to the respective speaker's pitch range in the latter case). Some of the stimuli were presented as video clips showing Smartakus speaking, but without correct lip synchronization. Participants were asked to rate the stimuli for naturalness on a five-point scale from -2 to +2, where -2 corresponded to "not natural" and +2 corresponded to "very natural". Mean scores were calculated for every stimulus.

The most important outcome of the evaluation procedure was that the subjective ranking of speakers was different for the two steps. For instance, the left panel in Figure 3 shows that for the best four male speakers from the first step, the MBROLA

[1] see http://tcts.fpms.ac.be/synthesis/

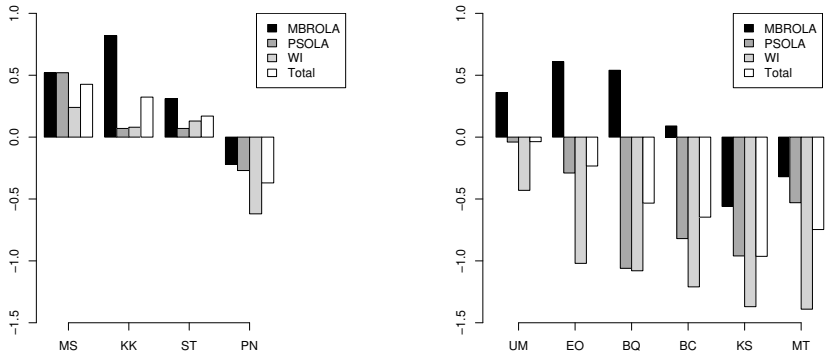


Fig. 3. Mean scores of audio stimuli for the male voices (left panel) and the female voices (right panel), broken down by signal processing method. Speakers are indicated on the x-axis, mean scores on the y-axis. Averaged over the different methods (white bars), the MS voice was rated the most natural, but when looking at MBROLA voices only (black bars), KK's voice was clearly better.

diphone voice of KK, who was originally ranked third, was judged to be the most natural diphone voice, and the second most natural diphone voice was from MS, who was originally ranked fourth. Other signal processing methods yielded different rankings; in these cases, the MS diphone voice was judged to be the most natural. Similar effects are evident in the ranking of the female diphone voices in the right panel of Figure 3. This confirmed our expectation that the subjective quality of the diphone voice does not correlate directly with the subjective quality of the original voice.

It is evident in Figure 3 that MBROLA turned out to be the best signal processing method in all cases. Male voices were generally rated better than female voices, especially for signal processing methods other than MBROLA.

In spite of Smartakus' relatively androgynous features, there was an even stronger preference for the male voices in the video clips. This is illustrated in Figure 4. Only the results for the MBROLA voices are presented here, which were again rated better than the other voices. In the video condition, only the speakers' original prosody was used, which was transplanted onto the diphone voices. To assess the influence of the natural prosody in the ratings of the video stimuli, we included the ratings for audio stimuli with natural prosody in the diagram. The preference for MS in the video condition was not due to the natural prosody: while MS and KK were rated similarly good for those stimuli, MS was clearly preferred in the video condition. Thus, as alluded to above, the speaker rated best for the audio-only stimuli was rated much lower for the audio-video stimuli, presumably because of his low pitch, which did not seem to go well with the cartoon-like features of Smartakus.

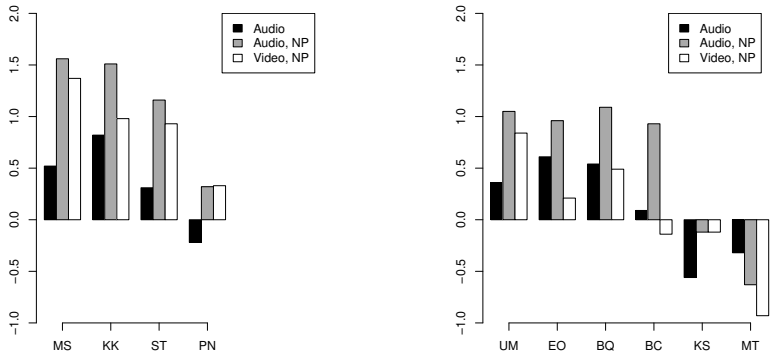


Fig. 4. Mean scores for MBROLA voices presented in audio mode, averaged over natural and rule-based prosody (black bars), with natural prosody (NP) only (gray bars), and for MBROLA voices presented in video mode with natural prosody only (white bars). All except one voice (MT) are rated better when the speaker's natural prosody is used in the audio-only condition. In the video condition, the ranking is different. MS is rated best in video mode. Generally, the male voices are preferred in this mode.

Based on these results, MS was selected as the speaker for SmartKom. We recorded both a diphone and a unit selection database of this speaker. The diphone voice was used in the project during the development of the unit selection voice, and it was also used as a baseline synthesis voice in the evaluation of the unit selection voice.

3 Restricted domain unit selection synthesis

The SmartKom domains are restricted but not limited: utterances are generated from a number of lexicalized partial syntactic trees [4] (see Chapter **??**), 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 [23]: 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 [27]. It is described in the following section. Details on the construction of the unit selection corpus are presented in Section 3.2.

3.1 Unit selection strategy

Current unit selection approaches mostly use segments [17, 7, 8] or subsegmental units such as half-phones [6, 11] or demiphones [3] as the basic unit. For each unit in the target utterance, several candidates are selected from the speech database according to criteria such as segment identity, segmental and linguistic context. For each candidate, its *target cost* expresses how well it matches the specification of the target unit. For each pair of candidates, their *concatenation cost* measures the acoustic distortion that their concatenation would cause. Then the sequence of candidates is chosen which simultaneously minimizes target and concatenation costs. Since there is no distortion for originally adjacent units, longer stretches of successive units are favored over single non-adjacent units, reducing the number of concatenation points and rendering a more natural voice quality. We will call this a bottom-up approach because, starting from the segmental level, the selection of complete syllables, words or phrases arises indirectly as a consequence of the lower concatenation costs for adjacent segments.

Such an approach faces two challenges. First, target costs and concatenation costs must be carefully balanced. Second, for frequent units the candidate sets can be very large, and the number of possible sequences of candidates grows dramatically with the number of candidates. For performance reasons, the candidate sets must be reduced, at the risk of excluding originally adjacent candidates.

One way to achieve the reduction of unit candidate sets is to cluster the units acoustically in an off-line procedure and to restrict the candidate set to the units of the appropriate cluster [7]. We will refer to this method as the acoustic clustering (AC) approach. The idea is to cluster 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. In other words, the linguistic properties that divide the units into acoustically similar clusters are those properties that apparently have the strongest influence on the acoustic realization of the units in the cluster. During synthesis, the linguistic context determines the pertinent cluster. All other units are ignored, which reduces the number of candidates.

Some approaches [32, 31] use a different strategy. 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. For the phonological structure matching (PSM)

algorithm [32], candidates can correspond to various nodes of the metrical tree of an utterance, ranging from phrase level to segment level, while [31] uses only the word and segment levels. Both approaches are designed for limited domains and benefit 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.

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 strategy that integrates the two approaches described above. 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.

Our implementation of the PSM algorithm differs from the original implementation [32] in some aspects. First, the original algorithm requires candidates to match the target specification with respect to tree structure and segment identities, but they may differ in stress pattern or intonation, phonetic or phrasal context, at the expense of higher target costs. This reflects the view that a prosodically suboptimal but coherent candidate is better than the concatenation of smaller non-coherent units from prosodically more appropriate contexts. We kept the matching condition more flexible by more generally defining two sets of features for each level of the linguistic hierarchy. *Primary features* are features in which candidates have to match the target specification (in addition to having the same structure), while they may differ in terms of *secondary features*. Mismatch of secondary features causes higher target costs, just as the mismatch of prosodic features increases the unit score in the original algorithm. The primary features typically are the unit identity and the classification of prosodic events occurring on the respective unit. Secondary features are mostly positional features expected to have a strong influence on the acoustic realization of the unit. More details can be found in [27].

Another, more important, difference to the original PSM algorithm is that candidate sets can optionally be reduced if their size exceeds a certain threshold. In this case, the candidate set is filtered stepwise for each secondary feature, thereby excluding candidates that do not agree on the respective feature, until the size of the candidate set is below the threshold. However, the PSM search is not performed below the syllable level because the initial candidate sets would be too large. Instead, the AC algorithm [7] takes over on the segment level, adding candidates for those parts of the target utterance that have not been covered yet.

As for the final selection of the optimal sequence of units, candidate units found by either search strategy are treated in the same way, i.e., they are subject to the same selection procedure. Thus, longer units are treated just as shorter units in that the optimal sequence of candidates is determined by a Viterbi algorithm, which si-

multaneously minimizes concatenation costs and target costs. Concatenation costs for two longer units are the concatenation costs for the two segments on either side of the concatenation point.

3.2 Text material design and corpus preparation

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 [31] 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 [33, 23].

We followed [33] by applying a greedy algorithm to select from a large text corpus a set of utterances which maximizes coverage of units. The procedure was as follows. First, the linguistic text analysis component of the IMS German Festival TTS system [18, 29] was used to determine for each sentence in a German newspaper corpus of 170 000 sentences the corresponding phone sequences as well as their prosodic properties. We built a vector for each segment including its phonemic identity, syllabic stress, word class, prosodic and positional properties. Thus, we obtained a sequence of vectors for each sentence. Additionally, we determined the diphone sequence for each sentence. 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, we constructed sentences that would contain them, added these sentences to the corpus, and repeated the selection process. 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 people's 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 using the same professional speaker as for the diphone voice and amounts to about 160 minutes of speech. The automatically generated transcriptions were manually corrected according to what the speaker had said together with the corresponding orthographic notation. The hand-corrected transcriptions were then used for sentence-wise forced alignment of the speech signal on the segment, syllable and word levels. Pitch accents and boundary tones were automatically predicted from the orthographic notation and subsequently corrected manually.

The corrected version of the database contains 2488 diphone types. 277 of the 2377 originally predicted types were not realized in the database, mostly because of incorrect predictions; instead, 388 additional types occurred. Similarly, the database had been predicted to cover 2731 out of 2932 phoneme/context vector types from the complete text corpus. 687 of these were not realized in the recorded database, whereas 791 new ones occurred, which yields 2835 types. Of these new vector types,

only 10 belong to the 201 vectors that had been in the complete text corpus but not in the subset selected for the recordings.

These figures show that more than 90% of the diphone types were covered as expected, and many new types involving foreign phonemes were added. As for the coverage of phoneme/context vectors, the situation is more complex. Combinatorially, 19 440 phoneme/context vector types are possible. We estimate that no more than 4600 are theoretically possible because the context properties are not independent. For instance, boundary tones only occur on phrase-final syllables. Some consonants are phonotactically not allowed in syllable onsets, others not in the rhyme, and vowels are in the rhyme per definition. Also, pitch accents are always realized on syllables with syllabic stress, and function words usually have no pitch accent. However, only approximately 60% of these 4600 types were covered even with a careful database design. One reason for this is that some of these types are so rare that they do not occur even in large corpora [23]. Apart from that, coverage of phoneme/context vectors was problematic because many of the predicted vectors were incorrect. This was partly due to foreign language material in the text corpus which could not be adequately dealt with using the monolingual German lexicon; also, unknown words, mostly compounds, abbreviations and acronyms, had often been predicted incorrectly. We expect that the prediction of context vectors can be significantly improved if foreign material is reliably marked as such in a preprocessing step. However, the prosodic contexts are difficult to predict, and often several alternative realizations are possible. Giving the speaker additional directions concerning intended prosodic realizations, on the other hand, may add too much load in supervising the recordings and moreover might result in unnatural realizations.

4 Lip and gesture synchronization

It has been shown that visual segmental information can enhance segmental speech perception [20]. Vice versa, inconsistencies between visual and acoustic information can significantly decrease intelligibility to the extreme that the segmental identity is compromised: [19] demonstrated that acoustic [ba] is perceived as /da/ when presented with the visual information of [ga]. Thus, correct lip synchronization is an important issue in multimodal speech synthesis.

4.1 Lip synchronization

In contrast to lip synchronization for more human-looking avatars which may require modeling of various parameters such as the position of the jaw, the upper and lower lip, teeth, tongue tip, and tongue root, only two parameters are necessary for Smartakus because of his cartoon-like features: jaw opening and lip rounding. His teeth and tongue are never visible.

We used a simple mapping procedure to map phonemes to so-called *visemes*. Visemes are visually contrastive speech elements [15]. In our view, visemes are sets of feature-value pairs, where the features correspond to the different articulators and

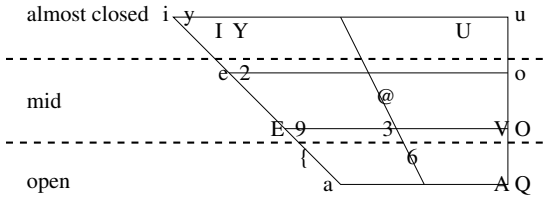


Fig. 5. IPA vowel space diagram of the vowels used in SmartKom synthesis, in SAMPA notation. The lip opening degree for each vowel is indicated on the left. The dashed lines indicate the boundaries between different degrees of opening.

the values indicate their target positions. Movement results from interpolating between the target positions specified by the visemes. Each phoneme is represented by one ore more visemes. Visemes can be underspecified regarding particular features. In this case, the value of the feature consists of a range of possible values. Underspecified visemes inherit the missing values from the context, which allows us to model coarticulation.

In SmartKom, only two features are specified for visemes: lip (or jaw) opening and lip rounding. Four degrees of lip opening (closed, almost closed, mid, open) and two degrees of lip rounding (unrounded, rounded) are differentiated. Visemes corresponding to **vowels** are fully specified for both opening and rounding. Figure 5 shows the vowels used in SmartKom arranged in the International Phonetic Association (IPA) vowel diagram, in SAMPA[2] notation. The position of each vowel in the IPA diagram reflects the tongue position in articulation. Thus, the vertical position of the vowels indicates tongue height, and the horizontal position indicates the front-back position of the tongue. To map vowel positions in the diagram to lip opening degrees, we stipulate that tongue height and lip opening correlate, but that the horizontal position is irrelevant for lip opening. The resulting mapping from the position in the diagram to the opening degree of the corresponding viseme is indicated on the left of Figure 5. Schwa vowels (/6/, /@/) are an exception: they are usually realized in a reduced way, in which case the correlation between lip opening and tongue height seems to be less strong. Both schwas are therefore realized with almost closed lips. Rounding trivially follows the phonological specification of the respective vowels. Diphthongs are represented by a sequence of two visemes corresponding to the visemes representing the underlying vowels.

Visemes for **consonants**, on the other hand, may be underspecified. This is motivated by the hypothesis that consonants whose place of articulation is to the back can be articulated with an almost closed jaw or with an open jaw. Consonants with an anterior place of articulation, however, can only be articulated with a relatively closed jaw. Also, most consonants can be articulated with rounded or with unrounded lips. Thus, visemes corresponding to consonants are unspecified with regard to lip rounding, and the farther back their place of articulation is, the higher is the degree of

[2] SAMPA is a wide-spread standard used for convenience instead of the IPA notation.

	place of articulation								
	labial	labiodental	dental	alveolar	postalveolar	palatal	velar	uvular	glottal
plosives	p b		t d				k g		?
nasals	m		n				N		
fricatives		f v	T D	s z	S Z	C	x	R	h
approximants			l r			j	w		
	jaw opening								
minimal	0	1	1				1		
maximal	0	1	2				3		

Fig. 6. IPA chart for the consonants used in SmartKom, in SAMPA notation (upper part of the table) and mapping to ranges of jaw opening degrees (lower part). Places of articulation are arranged in the chart from front (labial) to back (glottal). Possible jaw opening degrees are 0 (closed), 1 (almost closed), 2 (mid), and 3 (open). The range of values depends on the place of articulation of the corresponding phoneme. Velar, uvular, and glottal consonants exhibit the highest degree of underspecification for jaw opening: the range of jaw opening degrees is from 1 to 3.

underspecification for jaw opening. In Figure 6, the mapping from place of articulation to jaw opening degree is again demonstrated by an IPA chart containing the consonants used in SmartKom.

After mapping phonemes to visemes, the resulting sequences of partly underspecified visemes and the corresponding time intervals relative to the beginning of the speech signal are passed to the presentation manager (see Chapter **??**), which resolves underspecification and concatenates video sequences corresponding to the visemes.

4.2 Gesture-speech alignment

Smartakus may execute gestures while he is speaking. In this case, temporal alignment of speech and gesture is required. Pointing gestures also influence prosody, as mentioned in Section 1.

Building on the Sketch Model [13], speech is synthesized independently of the temporal structure of the accompanying gesture. Instead, the gesture is executed in temporal alignment with the speech signal. According to [13], gestures can be divided in three phases, viz. the preparation phase, the stroke phase, and the retraction phase. The stroke phase is the core phase, which accompanies corresponding speech material. Preparation and retraction phases of gestures can be adjusted to align the stroke phase with the relevant speech material. In SmartKom, most of the gestures occurring during speech are pointing gestures, which accompany deictic elements in the linguistic structure. In this case, the timing information for the deictic material is passed to the presentation manager to enable alignment of the stroke phase with the corresponding deictic element.

5 Evaluation of the speech synthesis module

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 when it comes to user acceptance or usability. Consequently, the SmartKom system has been evaluated extensively as a whole (Chapter **??**).

However, in addition to an end-to-end evaluation of the complete system, the evaluation of its 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 voice 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[3]. 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 with the diphone voice. Figure 7 shows an overview of the tasks performed in the evaluation procedures.

5.1 First evaluation

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 [5]) recorded from a speaker, and one using SUS stimuli synthesized by the diphone voice.

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

[3]The significant contributions of Martin Ernst (DaimlerChrysler, Ulm) and Gerhard Kremer, Wojciech Przystas, Kati Schweitzer, and Mateusz Wiacek (IMS) to the synthesis evaluations are gratefully acknowledged.

	1st evaluation		2nd evaluation			
			pilot study		full experiment	
	material	voice	material	voice	material	voice
dictation	SUS	natural			SUS	diphones
		diphones				US
	SK	diphones				
listening comprehension					open domain	diphones
						US
subjective impression	SK	diphones	SK	diphones	SK	US
				US		

Fig. 7. Overview of the tasks performed by participants in the evaluation procedures. The general type of task is indicated in the left column. The table lists text material, viz. normal text (open domain), semantically unpredictable sentences (SUS), or SmartKom specific material (SK), and the voices used to generate the stimuli, viz. diphone voice or unit selection voice (US).

typing errors, and in 2% of 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.

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 Figure 8. The material to fill the lexical slots in the templates came from lists of words selected from CELEX [2] according to their morphological and syntactic properties. The lists were randomized before generating the SUS stimuli. 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

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.)

Fig. 8. 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. 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.

/R/-diphones concatenated in these cases are two different positional variants of /R/, viz. a postconsonantal variant, and an intervocalic variant.

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, whereas only 72% of the naive subjects expressed the same opinion. In summary, the first evaluation confirmed that the diphone voice yielded satisfactory results.

5.2 Second Evaluation

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.

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).

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 (Section 3). 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.

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. Subjects were asked to answer three questions by adjusting a sliding bar between two extremes. The three questions were *"How do you judge the intelligibility of 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.

In the last task, 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 for intelligibility, naturalness, and pleasantness 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.

5.3 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.

References

1. Steven P. Abney. Chunks and dependencies: bringing processing evidence to bear on syntax. In *Computational Linguistics and the Foundations of Linguistic Theory*. CSLI, Stanford, 1995.
2. 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.
3. Marcello Balestri, Alberto Pacchiotti, Silvia Quazza, Pier Luigi Salza, and Stefano Sandri. Choose the best to modify the least: a new generation concatenative synthesis system. In *Proceedings of the European Conference on Speech Communication and Technology (Budapest, Hungary)*, volume 5, pages 2291–2294, 1999.

4. Tilman Becker. Fully lexicalized head-driven syntactic generation. In *Proceedings of the Ninth International Workshop on Natural Language Generation*, pages 208–217, Niagara-on-the-Lake, Ontario, Canada, 1998.
5. 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*, 18:381–392, 1996.
6. Mark Beutnagel, Mehryar Mohri, and Michael Riley. Rapid unit selection from a large speech corpus for concatenative speech synthesis. In *Proceedings of the European Conference on Speech Communication and Technology (Budapest, Hungary)*, volume 2, pages 607–610, 1999.
7. Alan W. Black and Paul Taylor. Automatically clustering similar units for unit selection in speech synthesis. In *Proceedings of the European Conference on Speech Communication and Technology (Rhodos, Greece)*, volume 2, pages 601–604, 1997.
8. Andrew P. Breen and Peter Jackson. Non-uniform unit selection and the similarity metric within BT's Laureate TTS system. In *Proceedings of the Third International Workshop on Speech Synthesis (Jenolan Caves, Australia)*, pages 373–376, 1998.
9. Daniel Büring. *The Meaning of Topic and Focus – The 59th Street Bridge Accent*. Routledge, London, 1997.
10. Guglielmo Cinque. A null theory of phrase and compound stress. *Linguistic Inquiry*, 24(2):239–297, 1993.
11. Alistair Conkie. Robust unit selection system for speech synthesis. In *Collected Papers of the 137th Meeting of the Acoustical Society of America and the 2nd Convention of the European Acoustics Association: Forum Acusticum (Berlin, Germany)*, 1999. Paper 1PSCB_10.
12. Peter W. Culicover and Michael S. Rochemont. Stress and focus in English. *Language*, 59(1):123–165, 1983.
13. Jan Peter de Ruiter. The production of gesture and speech. In David McNeill, editor, *Language and gesture*, pages 284–311. Cambridge University press, Cambridge, 2000.
14. Thierry Dutoit, Vincent Pagel, Nicolas Pierret, François Bataille, and Olivier van der Vrecken. The MBROLA project: Towards a set of high quality speech synthesizers free for use for non commercial purposes. In *Proceedings of the International Conference on Spoken Language Processing (Philadelphia)*, volume 3, pages 1393–1396, 1996.
15. C. G. Fisher. Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, 11:796–804, 1968.
16. Daniel J. Hirst. Detaching intonational phrases from syntactic structure. *Linguistic Inquiry*, 24:781–788, 1993.
17. Andrew J. Hunt and Alan W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of the IEEE International Conference on Acoustics and Speech Signal Processing (München, Germany)*, volume 1, pages 373–376, 1996.
18. IMS German Festival home page. [http://www.ims.uni-stuttgart.de/phonetik/synthesis/index.html], 2000.
19. John MacDonald and Harry McGurk. Visual influences on speech perception process. *Perception and Psychophysics*, 24:253–257, 1978.
20. Dominic W. Massaro. *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. MIT Press, Cambridge, MA, 1998.
21. Jörg Mayer. Transcribing German intonation – the Stuttgart system. Technical report, University of Stuttgart, 1995.
22. Jörg Mayer. Prosodische Merkmale von Diskursrelationen. *Linguistische Berichte*, 177:65–86, 1999.

23. Bernd Möbius. Rare events and closed domains: Two delicate concepts in speech synthesis. *International Journal of Speech Technology*, 6(1):57–71, 2003.
24. Gregor Möhler and Alistair Conkie. Parametric modeling of intonation using vector quantization. In *Proceedings of the Third International Workshop on Speech Synthesis (Jenolan Caves, Australia)*, pages 311–316, 1998.
25. Edmilson Morais, Paul Taylor, and Fábio Violaro. Concatenative text-to-speech synthesis based on prototype waveform interpolation (a time-frequency approach). In *Proceedings of the International Conference on Spoken Language Processing (Beijing)*, pages 387–390, 2000.
26. Eric Moulines and Francis Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9:453–467, 1990.
27. Antje Schweitzer, Norbert Braunschweiler, Tanja Klankert, Bernd Möbius, and Bettina Säuberlich. Restricted unlimited domain synthesis. In *Proceedings of the European Conference on Speech Communication and Technology (Geneva)*, pages 1321–1324, 2003.
28. Antje Schweitzer, Norbert Braunschweiler, and Edmilson Morais. Prosody generation in the SmartKom project. In *Proceedings of the Speech Prosody 2002 Conference (Aix-en-Provence)*, pages 639–642, 2002.
29. Antje Schweitzer and Martin Haase. Zwei Ansätze zur syntaxgesteuerten Prosodiegenerierung. In *Proceedings of the 5th Conference on Natural Language Processing – Konvens 2000 (Ilmenau, Germany)*, pages 197–202, 2000.
30. Elizabeth Selkirk. *Phonology and Syntax – The Relation Between Sound and Structure*. MIT Press, Cambridge, MA, 1984.
31. 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 Wolfgang Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, pages 519–534. Springer-Verlag, 2000.
32. Paul Taylor and Alan W. Black. Speech synthesis by phonological structure matching. In *Proceedings of the European Conference on Speech Communication and Technology (Budapest, Hungary)*, volume 2, pages 623–626, 1999.
33. Jan P. H. van Santen and Adam L. Buchsbaum. Methods for optimal text selection. In *Proceedings of the European Conference on Speech Communication and Technology (Rhodos, Greece)*, volume 2, pages 553–556, 1997.