Restricted Unlimited Domain Synthesis

Antje Schweitzer, Norbert Braunschweiler, Tanja Klankert,
Bernd Möbius, Bettina Säuberlich

Institute of Natural Language Processing
University of Stuttgart, Germany
antje.schweitzer@ims.uni-stuttgart.de

Abstract

This paper describes the hybrid unit selection strategy for restricted domain synthesis in the SmartKom dialog system. Restricted domains are characterized as being biased toward domain specific utterances while being unlimited in terms of vocabulary size. This entails that unit selection in restricted domains must deal with both domain specific and open-domain material. The strategy presented here combines the advantages of two existing unit selection approaches, motivated by the claim that the phonological structure matching approach is advantageous for domain specific parts of utterances, while the acoustic clustering algorithm is more appropriate for open-domain material. This dichotomy is also reflected in the speech database, which consists of a domain specific and an open-domain part. The text material for the open-domain part was constructed to optimize coverage of diphones and phonemes in different contexts.

1. Introduction

This paper describes the unit selection strategy in the German SmartKom project [1]. SmartKom is a multi-modal concept-to-speech dialog system which can handle requests in several domains. Information is presented graphically and acoustically by a life-like artificial character. The domains comprise cinema information, an electronic programming guide (EPG) accessing the TV program, tourist information, route planning, telephony and address book management. This scenario poses several requirements to the synthesis module, two of which will be discussed here. First, the domains are restricted but not limited. Utterances are generated from a number of lexicalized partial syntactic trees [2] but open slots are filled with names, proper nouns, movie titles etc. from dynamic external and internal databases, i.e., the vocabulary is unlimited, although it is biased toward domain specific material. Second, movie titles from the cinema and EPG domain involve extraordinarily many foreign words and necessitate the extension of the German phoneme inventory with English and French phonemes.

Addressing the issue of restricted domains, the predominance of domain specific material calls for a unit selection approach with a domain specific database to ensure optimal 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: 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 material pose different requirements to 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. It is described in section 2.

Regarding the extended phoneme inventory, English and French phonemes were mapped to their German counterparts or to similar German phonemes if possible. The remaining English and French phonemes had already been included in a previously recorded diphone database. For several reasons, the foreign phonemes could not be systematically covered in the unit selection corpus, requiring the existing diphone database to be kept as an additional database for unit selection if necessary. More details on the construction of the unit selection corpus are presented in section 3.

2. Unit selection strategy

Current unit selection approaches mostly use segments [3, 4, 5] or sub-segmental units such as half-phones [6, 7] or demiphones [8] 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 segment 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 exponentially 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 acoustically cluster the units in an off-line procedure and to restrict the candidate set to the units of the appropriate cluster

[4]. We will refer to this 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, reducing the number of candidates. Another advantage is the fact that this approach avoids specification of concrete acoustic properties of the target unit. Instead, the properties are implicit in the choice of the cluster depending on the linguistic context.

Some approaches [9, 10] 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 [9], candidates can correspond to various nodes of the metrical tree for an utterance, ranging from phrase to segment level, while [10] 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 is expected to ensure 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 [9] 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 patterns or intonation, phonetic or phrasal context, at the expense of a higher unit “score” (0 being the optimal score). 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 in 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 increased the unit score in the original algorithm.

An example for a target specification including primary and secondary features is given in figure 1. The relevant linguistic levels are (from top to bottom) the phrase, word, syllable, and segment level. On the phrase, word and syllable level, the primary features include phonetic transcription and stress as well

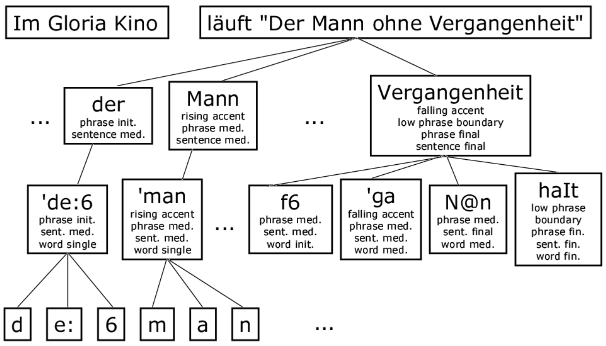

Figure 1: Target specification for (parts of) an utterance *(The Gloria theater shows “The Man Without A Past”)*. The linguistic levels are (from top to bottom) the phrase, word, syllable, and segment level. Values of the features are annotated on each node. Labels printed in large font correspond to primary features; secondary feature values are indicated by the smaller font size. For simplicity, the units on the phrase and word level are represented by their orthographic transcription instead of phonetic symbols.

as phrase, word and syllable boundaries. The secondary features differ on each level. Currently, there are no secondary features on the phrase level; on the word level, pitch accent type (falling vs. rising) and boundary tone (low vs. high), position of the word in the phrase (initial, medial or final for phrases containing more than one word, or single for phrases containing a single word) and position of the word in the sentence (initial, medial, final or single) are considered. Agreement on these features is also required on the syllable level, with the addition of the position of the syllable in the word (initial, medial, final or single).

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, 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 [4] 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 the same, 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 simultaneously minimizing 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. 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 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 [11, 12].

We followed [11] in 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 prosody prediction component of the IMS German Festival text-to-speech system [13, 14] was used to determine for each sentence in a German newspaper corpus of 170 000 sentences the phone sequences as well as the prosodic properties. We built a vector for each segment including (i) its phonemic identity, (ii) its position in the syllable (onset or rhyme), (iii) presence or absence of syllabic stress on the corresponding syllable, (iv) type or absence of pitch accent on the syllable, (v) type or absence of boundary tone on the syllable, (vi) position of the syllable in the phrase (initial, medial and final) and (vii) word class of the related word (function word or content word). Thus, we obtained a sequence of vectors for each sentence. Note that these prosodic properties partly correspond to the primary features used for the selection of units (cf. section 2). 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 full diphone coverage could be obtained, and at the same time the number of phoneme/context vector types was increased.

We faced several difficulties with this procedure. For instance, the automatic transcriptions of the sentences contained many errors, which were caused by the high proportion of words or expressions not contained in the lexicon. These included German compound nouns, unknown abbreviations, and foreign words, mostly English. The latter constituted a major problem because we had no means to recognize them as foreign, therefore automatically applying German letter-to-sound rules to them. Apart from the incorrectness of the resulting transcriptions, this also prevented systematic coverage of English phonemes. Instead we had to rely on the domain specific part with its many English movie titles for an appropriate coverage of these phonemes, and on the existing diphone database as a backup. Another challenge was that incorrect transcriptions often involved unusual phoneme combinations and the corresponding sentences were all the more prone to be selected by the greedy algorithm. To alleviate this problem, we excluded obvious transcription errors prior to selection partly automatically and partly manually, leaving approximately 100 000 sentences for the algorithm to choose from.

The statistics of the selected subset of sentences is as follows. We truncated the list of selected sentences at 1557, which was the minimal set of sentences predicted to cover the complete set of 2377 mostly German diphone types as well as 2731 out of 2932 phoneme/context vectors in the corpus altogether. The most frequent vector in our sub-corpus occurred 2287 times, representing the segment /@/ in a content word in phrase medial position with no pitch accent and no boundary tone. Other phonotactically possible but very unlikely vectors did not occur at all, such as the segment /9/ in a stressed syllable of a function word in phrase-final position without pitch accent but a high boundary tone. We did not manually construct sentences for the missing vector types.

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 latter was necessary because acronyms, abbreviations and numbers had often been incorrectly expanded. The hand-corrected transcriptions were then used for sentence-wise forced alignment of the speech signal on the segment, syllable and word level. The orthographic notations were used to correct the predicted prosodic contexts. We automatically predicted pitch accents and boundary tones and generated label files for them.

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 them were not realized in the recorded database, while 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.

4. Results and discussion

Although we have not carried out a formal evaluation yet, it is evident that the subjective quality of the unit selection voice by far exceeds the quality of the diphone voice. Preliminary tests with 31 distinct utterances from realistic test runs with the dialog system confirm that our algorithm succeeds in selecting relatively long sequences of successive units: candidate units found by our PSM implementation had a mean length of 5.5 segments, with a maximum of 46 segments, for the phrase *Herzlich willkommen beim SmartKom-Informationssystem ("Welcome to the SmartKom information system")*, which is part of the introductory system turn. After concatenation, the mean length of coherent sequences was 6.0 segments. Audio files for these 31 synthesized utterances can be found at *http://www.ims.uni-stuttgart.de/phonetik/unitselection*. A formal evaluation of the unit selection voice is about to be prepared at the time of writing this paper. Before performing this evaluation, the segment labels and the prosodic labels in the database will be manually corrected to ensure that synthesis quality is not affected by label errors.

An open issue concerns the primary and secondary features for the PSM candidate search. It has not been finally decided yet which features should be primary features, in which the candidates must exactly match the target specification. Using few primary features causes more candidates to be found, but, on the other hand, will entail that the search for candidates is not continued on lower levels of the linguistic representation. We suggest that systematic variation of primary and secondary features and the evaluation of the respective results is a sensible procedure for finding the optimal demarcation between primary and secondary features. Related to this issue, the question arises whether the secondary features should be weighted individually when determining the target costs. Individual weights could be found by systematic variation as well.

Regarding the database design, more than 90 per cent 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 not 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 per cent 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 copora. 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 may add too much load in supervising the recordings and moreover could result in unnatural realizations.

5. Conclusion

We have presented a hybrid approach for unit selection synthesis in restricted domains. Restricted domains are characterized as being biased toward domain specific material, while being unlimited in that they involve a substantial amount of open-domain material. We have argued that both kinds of material pose different requirements to the selection strategy and to the speech database. Regarding the selection strategy, two existing approaches, viz. the phonological structure matching and the acoustical clustering approach, have been integrated to deal with both kinds of material in a uniform way. As for the speech database, we have proposed to combine domain specific material with a carefully constructed open-domain database that was optimized with respect to coverage of diphone types and phonemes in different prosodic contexts. We have presented details on the corpus construction process and on our implementation of the selection algorithm. Preliminary results show that the synthesis quality by far exceeds the quality of the existing diphone voice of the same speaker, particularly with respect to naturalness and voice quality.

6. Acknowledgments

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.

7. References

[1] W. Wahlster, N. Reithinger, and A. Blocher, “SmartKom: Multimodal communication with a life-like character,” in *Proceedings of the European Conference on Speech Communication and Technology (Aalborg, Denmark)*, vol. 3, 2001, pp. 1547–1550.

[2] T. Becker, “Fully lexicalized head-driven syntactic generation,” in *Proceedings of the Ninth International Workshop on Natural Language Generation*, Niagara-on-the-Lake, Ontario, Canada, 1998, pp. 208–217.

[3] A. J. Hunt and A. 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)*, vol. 1, 1996, pp. 373–376.

[4] A. W. Black and P. Taylor, “Automatically clustering similar units for unit selection in speech synthesis,” in *Proceedings of the European Conference on Speech Communication and Technology (Rhodos, Greece)*, vol. 2, 1997, pp. 601–604.

[5] A. P. Breen and P. 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)*, 1998, pp. 373–376.

[6] M. Beutnagel, M. Mohri, and M. 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)*, vol. 2, 1999, pp. 607–610.

[7] A. 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.

[8] M. Balestri, A. Pacchiotti, S. Quazza, P. L. Salza, and S. 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)*, vol. 5, 1999, pp. 2291–2294.

[9] P. Taylor and A. W. Black, “Speech synthesis by phonological structure matching,” in *Proceedings of the European Conference on Speech Communication and Technology (Budapest, Hungary)*, vol. 2, 1999, pp. 623–626.

[10] K. Stöber et al., “Speech synthesis using multilevel selection and concatenation of units from large speech corpora,” in *Verbmobil: Foundations of Speech-to-Speech Translation*, W. Wahlster, Ed. Springer-Verlag, 2000, pp. 519–534.

[11] J. P. H. van Santen and A. L. Buchsbaum, “Methods for optimal text selection,” in *Proceedings of the European Conference on Speech Communication and Technology (Rhodos, Greece)*, vol. 2, 1997, pp. 553–556.

[12] B. 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.

[13] “IMS German festival home page,” [http://www.ims.uni-stuttgart.de/phonetik/synthesis/index.html].

[14] A. Schweitzer and M. Haase, “Zwei Ansätze zur syntaxgesteuerten Prosodiegenerierung,” in *Proceedings of the 5th Conference on Natural Language Processing – Konvens 2000 (Ilmenau, Germany)*, 2000, pp. 197–202.