Unit Selection Synthesis in the SmartWeb Project

Martin Barbisch, Grzegorz Dogil, Bernd Möbius, Bettina Säuberlich, Antje Schweitzer

Institute for Natural Language Processing
University of Stuttgart, Germany
{Firstname.Lastname}@ims.uni-stuttgart.de

Abstract

This paper describes three aspects of the unit selection synthesis used in the SmartWeb dialog system. The synthesis module has been implemented in the IMS German Festival speech synthesis system. First, we compare a unit selection strategy developed in the course of the project to a strategy developed earlier. Second, we discuss our experiences with F0 smoothing and amplitude modeling, which were both devised to reduce audible discontinuities. However, the results are inconclusive so far. Finally, we sketch a simple mechanism that addresses the problem of language disambiguation for proper names.

1. Introduction

SmartWeb is a research project funded by the German government [1]. The goal of the project is to implement a mobile intuitive user interface to the *Semantic Web* which allows requests involving natural speech and gestures. Answers are also rendered by speech, which is synthesized by the unit selection synthesis module described in this paper.

The synthesis module used in SmartWeb is based on the synthesis module developed in the predecessor project [2] and is implemented in the IMS German Festival framework [3].

In the course of the SmartWeb project, we have built two databases, one for a male speaker, and one for a female speaker. We have added a new unit selection strategy as an alternative to the existing strategy. Thus, there are two different unit selection algorithms available using the same database, text preprocessing and symbolic synthesis components. Both variants render very natural and intelligible speech. We compared the two variants in a first perception experiment to verify the validity of the new approach. The two variants and the perception experiment are described in some detail in section 2.

Although the synthesis results are very good altogether, there are some occasional glitches that seem to be caused by discontinuities in amplitude and pitch. We therefore experimented with amplitude modeling and different F0 discontinuity penalties. However, the results are inconclusive so far. The experiments and their results are discussed in section 3.

One key application of SmartWeb is the access to information on the soccer World Championships 2006. In this scenario, we faced the problem that proper names, particularly first names, are often ambiguous between several languages. We briefly sketch a simple mechanism to deal with this problem in section 4.

2. Comparing the two unit selection approaches

Both approaches combine aspects of two existing unit selection approaches, viz. phonological structure matching (PSM, [4]) and acoustic unit clustering (AC, [5]). We will call the first approach PSM/AC in the following because it combines PSM and AC in a straightforward way. The alternative approach will be called PSM/MC because in contrast to the original AC, the clustering is carried out manually.

2.1. PSM versus AC

The PSM algorithm [4] employs a top-down strategy for selecting the units from a speech database in which all sentences are represented as phonological tree structures. For each target sentence to be synthesized, the corresponding target tree structure is calculated. The PSM algorithm starts on the sentence level by comparing the available sentence tree structures to the target tree structure and possibly descends in the target tree structure until matching candidates are found. Generally, on any level a candidate matches if the trees below the target node match. If no adequate candidate is found on one level, the algorithm descends to the next lower level by assigning the daughters of the current node as new targets. This approach ensures that the longest available unit from the database is selected, minimizing concatenation points.

By contrast, the AC algorithm [5] only searches for candidates on the segment level. Longer continuous stretches of speech are only favored indirectly because they cause no concatenation costs later on. As the number of candidates is usually very high on the segment level, the candidates are clustered in an offline process. This is done automatically by creating a decision tree for each phoneme type with its leaves representing clusters of similar items. The features that are used for the questions at the nodes of the decision tree are linguistic-phonological features. The trees are built in a way that the acoustic similarity within the cluster is maximized, selecting only features that are significant in partitioning the tree. Thus, in building the tree, those features are determined that have the greatest impact on the acoustic realization.

The clusters can be pruned in order to obtain smaller clusters, by excluding segments that are farthest from the center of the cluster. This is intended to remove potentially poorly articulated or incorrectly labeled units. A second type of pruning is aimed at reducing units that are very common by removing units that are very similar to other existing units.

During the synthesis process, for each target segment the relevant cluster is determined by selecting the cluster which matches the desired linguistic-phonological context. The units belonging to that cluster are then taken as candidates.

The disadvantage of the AC algorithm is that in some cases the selection of continuous segments from the database is prohibited because they have either been assigned to a cluster which is not taken into consideration in the actual context, or because they have been removed during the pruning process.

Also during the construction of the decision trees no explicit linguistic and phonetic knowlegde is applied. For instance, it is impossible to give a higher priority to certain features, such as the manner and place of articulation of the context segments, which is expected to determine the strength and type of coarticulation effects.

The PSM algorithm on the other hand is problematic in open-domain scenarios because it does not restrict the number of candidates for each target unit. Particularly in open-domain scenarios, it will often be necessary to concatenate segment-level candidates because the database can not be tailored to cover all possible utterances by higher-level units. Since there will be very many segment candidates at least for the more frequent phonemes, the candidate network grows very large, reducing the efficiency of the algorithm.[1]

2.2. The PSM/AC approach

In the predecessor project to SmartWeb, we implemented a unit selection strategy combining PSM and AC by incorporating the strengths of both algorithms while avoiding their drawbacks in open-domain scenarios discussed above [2]. We call this approach the PSM/AC approach. The combination was motivated by the claim that PSM would prefer longer units in a more direct way than the AC approach, while clustering is appropriate to reduce candidate sets in cases where no long units are available.

Accordingly, PSM is used for phrase, word and syllable-sized units. If no appropriate candidates are available on the phrase, word, or syllable levels, AC is used on the segment level to reduce the segment candidate sets. This procedure ensures that at least longer units can be selected in their entirety; we do not run the risk that single segments within these units are not accessible because they have been assigned to another cluster or because they have been pruned during the clustering process.

2.3. The PSM/MC approach

The alternatively developed approach also employs the PSM strategy for candidate selection, but uses manual clustering (MC) to reduce the candidate sets on all levels, hence the name PSM/MC. The clustering is achieved by manually constructed decision trees. The use of decision trees on all unit levels allows for the consistent administration of all units and an efficient access via indexing.

The structure of the decision trees is given manually by ranking the features according to their linguistic-phonological relevance. The order of the ranking determines the questions at the nodes at each level of the decision trees. Each level of the tree represents a specific feature (e.g. place of articulation of the preceding or the next segment, or syllable stress, syllable position, etc.). The place of articulation of the segmental context is ranked very high in the decision tree as it is very important to model coarticulation effects.

The MC approach is highly flexible in that the decision tree can be easily reconstructed if a specific feature order turns out to be suboptimal. If no or only few candidates are found on a specific level, it is possible to collect all subordinate candidates on a higher level. Also the basic unit type can be selected freely (i.e. phone, diphone or demi-phone) with no further modification to the algorithm. This facilitates the comparison of the different basic unit types.

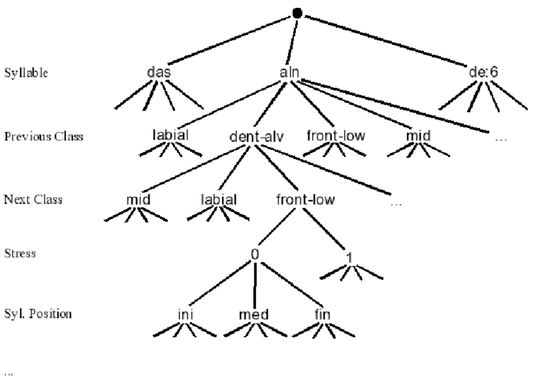

Figure 1: *Syllable level decision tree with exemplary feature ranking (left column). For each syllable type, this tree splits the candidates according to the place of articulation of the preceding and following segments (features "previous class" and "next class", respectively). Candidates are classified further according to the stress level of the syllable (feature "stress") and the position of the syllable in the phrase (feature "syl. position").*

The PSM/MC algorithm offers some advantages over PSM/AC. Firstly, its flexibility allows for comparing not only different basic unit types but also which phonetic features are most important for perception. The latter can be achieved by specifying different feature rankings for the phonetic features in question and rebuilding the decision trees with the respective order. This step requires no further manual interaction beyond the specification of the ranking and can easily be executed several times to test different rankings.

Secondly, PSM/MC usually does not run the risk of excluding or involuntarily ignoring potentially good candidates even before the selection process. Depending on the number of candidates, the selection process can be terminated on any level in the tree, selecting all candidates in the sub-trees beneath.

Thirdly, the clustering is adapted to the specific unit type and its requirements. This way phonetic knowledge can be directly applied in creating the decision trees. For instance, the place of articulation of the preceding segment is the primary selection criterion for all unit types. This is intended to model coarticulation effects, such as the influence of preceding labial consonants on the spectral properties of a vowel for instance, which would be expected to be different enough from the influence of, say, a preceding velar consonant to warrant the assignment of segments in these contexts to different clusters. On the syllable level, stress and the syllable's position in the corresponding phrase are important features. The high ranking of syllabic stress is motivated by the fact that it has been claimed to affect the spectral balance of the corresponding vowels [6, 7]. The position of the syllable in the phrase is expected to have an impact on the duration of the syllable and its segments as well as on their pitch level. Also, phrase-final segments and syllables are often laryngealized in German.

The disadvantage of PSM/MC lies in the statistically unbalanced distribution of the feature vectors in the corpus due to the

[1] For instance, on the segment level, our database contains 107,000 tokens representing 84 different German and foreign phonemes, which corresponds to an average of approx. 1,300 tokens per type, whereas on the syllable level, 41,000 tokens represent 3,350 syllable types, corresponding to an average of only 12 tokens per type.

LNRE characteristics of natural language [8], resulting in unbalanced decision trees. Since a few candidates are represented above average in the database, the trees exhibit large differences in the number of candidates at the leaf level, as units with identical feature vectors can not be differentiated and thus end up in identical clusters. Possible acoustic differences are not taken into account because MC only operates on the symbolic level, in contrast to AC, where the classification is driven by the signal.

2.4. Evaluation of PSM/AC versus PSM/MC

We compared both selection strategies in a first unsupervised perception experiment. We used the diphone-based version of PSM/MC because it was expected to model coarticulation effects better than the segment-based version. In this experiment, 26 subjects listened to 30 pairs of stimuli. The stimuli were 15 moderately long sentences (4 to 11 words) randomly selected from different text genres, synthesized using the two different algorithms and presented pairwise in different orders. Each pair could be played several times, but always in the same order. Listeners had to judge which stimulus sounded better, or if both stimuli sounded equally good, and they could take as long as they wanted to make their decision. In 22 cases, the stimuli in a pair were different (AB or BA order), and in 8 cases, they were identical. Participants were instructed that some stimuli would be identical. The identical pairs were included to assess the listeners' reliability.

Listeners favored PSM/AC over PSM/MC (49.8% vs. 40.7%, 9.4% undecided). The differences were statistically significant (χ^2(2,N=572)=154.03, p$\ll$0.05). The difference was not due to personal preferences, since only 3 participants consistently favored PSM/AC over PSM/MC (p<0.002)[2]. Instead, the differences were dependent on the stimulus pair: for 10 out of 22 pairs PSM/AC was rated significantly better, and for 6 of these pairs PSM/MC was rated significantly better (p<0.002)[3]. This means that for the majority of stimulus pairs, participants agreed in their judgment – they usually favored the same variant. The reason for this is that in some cases, the units selected from the database were not ideal realizations of the target unit, and that sometimes, the concatenation was suboptimal. These problems, which are typical for any unit selection algorithm, in some cases occurred in the PSM/AC stimulus, and in some cases in the PSM/MC stimulus, but the PSM/MC variant was affected slightly more often.

Altogether we consider the results of the evaluation encouraging enough to pursue the PSM/MC algorithm further, even more so because there are at least two aspects in which we are confident to improve the algorithm in the future.

First of all, an informal assessment of the specific problems in the PSM/MC stimuli suggests that the concatenation of diphones containing plosives was problematic in some cases, in that the corresponding stop releases could not be perceived properly. This is because our variant of the original optimal coupling algorithm [9] has been adapted to concatenate diphones by starting the search for a good concatenation point at the middle of the phoneme. The middle of the phoneme in case of stops is often close to the burst, and thus it happens occasionally that the burst is completely omitted when concatenating stops. One way to remedy this problem is to label the bursts in the database and to take the place of the burst into consideration when searching for the optimal concatenation point. Another way may be to modify the optimal coupling algorithm to detect the silence part of stops automatically. Compared to the first solution, this would eliminate the necessity to prepare the database beforehand.

Some additional improvement of the PSM/MC algorithm could be achieved by re-assessing the manually defined feature order in the selection trees. Although the current order was partly determined on the basis of phonetic knowledge, in some cases the ranking was not obvious and only preliminarily established by ad hoc decisions. These decisions might be reconsidered with the help of further perceptual evaluation procedures.

3. Prosodic modifications

In order to improve synthesis quality even further we investigated several possibilities for prosodic modifications. The motivation was that audible discontinuities seemed to be mostly caused by concatenation of prosodically too different candidates. Concerning pitch we experimented with different weights for the concatenation costs caused by pitch discontinuities. As for amplitude, we built a loudness model for each phoneme and adjusted actually selected segment candidates to fit those models.

3.1. Pitch Continuity

A smooth pitch contour is most important for intonation. Discontinuities of the pitch contour at unit boundaries cause audible glitches. In a first step, we investigated the influence of different weights for F0 differences in determining the concatenation costs.

The difference in F0 between consecutive units is already taken into account when calculating concatenation costs in Festival [9]. An additional weight factor has been added [10] to bring the costs caused by F0 differences into the same order of magnitude as the costs for spectral discontinuities. This weight factor has been predetermined for both the male and the female voice by synthesizing a large number of sentences and comparing the means for the spectral costs to the means for the F0 costs. The weight factor was chosen in a way that the same means are obtained for spectral costs and F0 costs. A second factor was defined in a configuration file which is intended to allow experimenting with different weight factors to give more or less priority to F0 continuity [10].

The effectiveness and usefulness of the newly introduced F0 weights were tested with three objective evaluation methods, varying the configurable weight factor to be 0, 1, 2, 3, or 5.

The first method was to compare the resulting F0 values with the idealistic F0 curve as predicted by the PaIntE model [11] by calculating the size of the area between the two curves:

$$curve_{RMSE} = \sqrt{\frac{\sum_{i=0}^{length(wave)} (f0(i) - f0_{PaIntE}(i))^2}{length(wave)}} \quad (1)$$

The smaller the area the better the F0 curve approximates the "optimum". However, the significance of this calculation depends on the quality of the reference curve and does not directly measure the smoothness of the F0 curve.

The second method was to determine an F0 curve "smoothness" correlate. The smoothness correlate was obtained by simply adding the absolute differences of consecutive F0 values in

[2] The significance level was adapted to p = 0.05/26 $\simeq$ 0.002 because of the 26 repeated χ^2 tests

[3] The significance level was adapted to p = 0.05/22 $\simeq$ 0.002 because of the 22 repeated χ^2 tests

the synthesized signal:

$$\sum_{i=0}^{n-1} |x_{i+1} - x_i| \qquad (2)$$

(where n is the number of frames and x the F0 value). If the smoothness increases with larger F0 weights, this is a good sign for fewer discontinuities in the F0 curve. However, this approach did not seem to be a good benchmark for the F0 cost function, since both increasing and decreasing smoothness were found for larger F0 weights, depending on the sentence synthesized.

The third method was to verify that different F0 weights did indeed have an effect in that they resulted in different candidates being selected, and to quantify the change by determining in how many cases different candidates were selected. The results confirmed that with increasing F0 weights, the number of different candidates increases as well. This was not generally the case; for some sentences, effects occurred only for F0 weight 3 or 5, while for others, there were changes even for an F0 weight of 1. This shows that the introduction of the additional weight factor successfully brings the F0 weights in an order of magnitude that is comparable to the weights applied to spectral differences. However, this does not answer the question whether the changes are positive or negative.

We conclude that a perceptual experiment similar to the one described in section 2.4 would be better suited to verify the usefulness of manipulating the F0 weight in the concatenation cost function.

3.2. Amplitude Modeling

Even for very carefully recorded speech databases, different realizations of one phoneme will have different sound levels, since they were produced in different contexts. Since loudness is no explicit selection criterion and only is taken into account when calculating the concatenation costs, it is possible that a unit is selected which fits perfectly except for the volume. To remedy this problem, we tried to apply, as the final step in synthesis, amplitude modification based on models we built before. The models were created by inspecting every occurrence of each phoneme in our database, measuring the RMSE values at 10, 25, 50, 75 and 90% of the phoneme duration and calculating the means [10]. In applying these models to the synthesized signal in the final step, each sample is multiplied by the factor determined by these models. Values between the calculation points are linearly interpolated. The procedure was based on the one used in the Bell Labs speech synthesis system [12, p. 222].

Pauses and plosives are not modified; the former since they have no energy, and the latter since they are hard to normalize due to their different phases (pause, burst and friction).

Figure 2 shows an example comparing the amplitude profile of the unmodified signal (blue dashed line) and the profile of the amplitude normalized signal (red solid line) for the phrase *einer der zentralen Plätze ("one of the central squares")*. The speech signal looks more natural after the modification. For instance, the [a:] is louder than the schwa [@] after the modification, which seems more natural than the other way round, which it was before the modification.

However, a perception experiment with 35 subjects using the same experimental procedure as described in section 2.4 showed that the unmodified signal was very clearly preferred over the amplitude normalized signal. The original signal was rated better for 52.2% of the stimulus pairs, while the normalized variant was preferred for only 12.0% of the pairs, and both variants were rated equally good for the remaining 35.8% of the pairs. The differences were statistically significant (χ^2(2,N=1040)=261.56,p$\ll$0.05). The accordance in listeners' judgements was overwhelming. Not a single listener consistently preferred the normalized variant. On the contrary, every listener rated the original variant better more often than the normalized variant, and this preference was significant for 15 out of 35 listeners (15 out of 35 listener-specific χ^2 tests yield values of p $< 0.05/35 \simeq 0.001$, and only 3 tests yield p > 0.05). Also, for no pair of stimuli, the normalized variant was rated better by more listeners than the original variant. The preference again was significant for most stimulus pairs (15 out of 23 stimuli-specific χ^2 tests yield values of p $< 0.05/23 \simeq 0.002$).

Given the negative outcome of the perception experiment, we analyzed the stimuli once again. On first visual inspection, the amplitude normalized variants seemed to be superior to the original variants. The amplitude profiles looked smoother and more natural, and usually, the normalization did not "stick out" perceptually. However, in few cases, the normalization caused problems for some segments. This occurred mostly for segments which exhibited lower amplitudes than expected. In these cases, the normalization resulted in boosting the respective segment too much, revealing phenomena that would otherwise not have been heard so clearly. In one example, a very low [l] segment contained an almost inaudible burst caused by the articulatory movement from a preceding [S]. After normalizing the [l] segment to the average amplitude of /l/ phonemes, the burst is perceived as an irritating noise. In another example, a problematic concatenation in a very low [@] segment caused a discontinuity that became much more obvious after the normalization. Some phonemes were generally problematic. For instance, syllable-initial vowels are often glottalized to some degree in German, there may even be the release of a glottal stop at the beginning of the vowel. These glottalized realizations have lower amplitudes than the non-glottalized modal-voiced realizations, and in these cases, raising the amplitude results in unnaturally loud glottal stops or glottalized vowel phases. Also, initial /h/ was generally boosted too much, making it sound almost like /x/. Thus, although the normalization for most segments did not compromise the quality of the speech signal, there was often at least one of the few problematic segments in the test sentences, and listeners seldom failed to detect them. A possible solution to this problem may be to limit the degree to which very low segments are manipulated, e.g. by assuming an upper limit for the normalization factor, but this will have to be investigated in the future.

4. Language disambiguation for proper names

In the course of the project, we added a simple language disambiguation component for proper names. Apart from the fact that proper names pose problems because of their often irregular pronunciation, particularly first names are often ambiguous between several languages. For instance, the first name *David* is pronounced differently depending on whether it is a German, English, French or Spanish name. However, the context often helps to disambiguate, e.g. in the above example, if the name *Beckham* follows, the English variant is obviously correct.

We have added a mechanism that facilitates disambiguation in such cases. This mechanism presupposes that there is a lexicon that not only contains transcriptions of the proper names

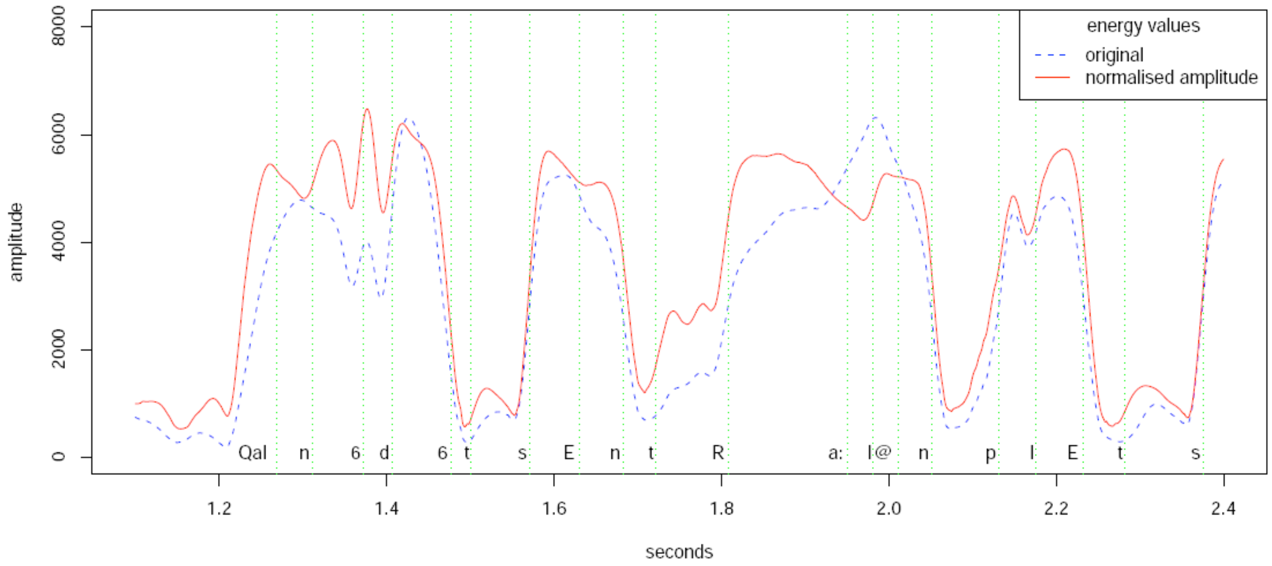

Figure 2: *Amplitude normalization. Note that the* [*a:*] *(1.8s) is louder than the* [*@*] *after the modification.*

but also that their origin is coded into the part of speech tag. In the example cited above, the lexicon would have to contain several entries for *David*, marked as a proper name of German, English, French, or Spanish origin, respectively, and one entry for *Beckham*, marked as English. This information could result from extracting proper names from foreign lexicons. In our experience, it is sufficient to differentiate between several languages or groups of languages. For instance, given the degree to which we adapted foreign pronunciations to our phoneme inventory, it was adequate to differentiate between English, German, and French, but the Spanish languages and Portuguese could be grouped together, and there was no further distinction necessary between different Slavic languages, or between different Asian languages.

The mechanism automatically collects all different transcriptions of orthographically identical proper names including their tags into a table that lists all possible origins for all ambiguous proper names. During synthesis, upon encountering an ambiguous name, the pronunciation is left underspecified by assigning the set of all possible origins to each name. Then, they are disambiguated by unifying the sets of possible origins of consecutive proper names.

Although we have only a moderate number of proper names that are marked for their origin (approximately 2,000 names), the mechanism has greatly improved the subjective synthesis quality because some of the most frequent cases of ambiguous names occurred very frequently in a SmartWeb key application, viz. the access of information on the soccer World Championships 2006.

5. Conclusions

We have described three aspects of the unit selection synthesis used in the SmartWeb system. First, we have described and compared two unit selection strategies. With respect to the PSM/MC strategy, the optimal feature rankings and the most adequate basic unit type should be investigated further. Particularly the optimal feature ranking will give interesting insights in the perceptual relevance of the respective features from a theoretical perspective. Another open issue is the treatment of bursts in concatenation, which should be addressed in the future. With these improvements, we expect the PSM/MC approach to surpass the PSM/AC approach in the future. For the time being, the PSM/AC approach is preferred over the PSM/MC approach in the SmartWeb project.

Second, we have discussed our experiences with different weights to enforce pitch continuity and with amplitude modeling. In the case of pitch continuity, we introduced an additional F0 weight factor that successfully brings the F0 weights in an order of magnitude comparable to the weights applied to spectral differences. However, a perceptual experiment to confirm the usefulness of manipulating the F0 weights has yet to be conducted. With regard to amplitude modeling, we found that it is clearly not useful, at least not in the way it has been applied here. Assuming an upper limit for the normalization factor may be expedient, but this has not been verified yet.

Finally, we have sketched a simple mechanism for language disambiguation of proper names that improved the subjective synthesis quality particularly for a SmartWeb key application.

6. Acknowledgments

This work was funded by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the SmartWeb project under Grant 01IMD01H. The responsibility for the content lies with the authors.

The perception experiment and the amplitude modeling were implemented by Andreas Madsack in the context of his diploma thesis. He also experimented with the pitch weights. The perception tests concerning the two selection strategies were carried out by Christian Scheible and Sylvia Riester.

7. References

[1] W. Wahlster, “Smartweb: Mobile applications of the semantic web,” in *KI 2004: Advances in Artificial Intel-*

ligence, S. Biundo, T. Frühwirth, and G. Palm, Eds. Berlin/Heidelberg: Springer, 2004, pp. 50 – 51.

[2] A. Schweitzer, N. Braunschweiler, G. Dogil, T. Klankert, B. Möbius, G. Möhler, E. Morais, B. Säuberlich, and M. Thomae, "Multimodal speech synthesis," in *SmartKom: Foundations of Multimodal Dialogue Systems*, W. Wahlster, Ed. Springer, 2004, pp. 411–435.

[3] "IMS German Festival home page," [http://www.ims.uni-stuttgart.de/phonetik/synthesis/], 2007.

[4] P. Taylor and A. W. Black, "Speech synthesis by phonological structure matching," in *Proceedings of the 6th European Conference on Speech Communication and Technology (Budapest, Hungary)*, vol. 2, 1999, pp. 623–626.

[5] A. W. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *Proceedings of the 5th European Conference on Speech Communication and Technology (Rhodos, Greece)*, vol. 2, 1997, pp. 601–604.

[6] K. Claßen, G. Dogil, M. Jessen, K. Marasek, and W. Wokurek, "Stimmqualität und Wortbetonung im Deutschen," *Linguistische Berichte*, vol. 174, pp. 202–245, 1998.

[7] M. Jessen, K. Marasek, K. Schneider, and K. Claßen, "Acoustic correlates of word stress and the tense/lax opposition in the vowel system of German," in *Proceedings of the 13th International Congress of Phonetic Sciences (Stockholm, Sweden)*, vol. 4, 1995, pp. 428–431.

[8] 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.

[9] "The Festival Speech Synthesis System," [http://www.cstr.ed.ac.uk/projects/festival/], 2007.

[10] A. Madsack, "Amplitude normalisation and intonation continuity modelling for unit selection," Diplomarbeit, IMS, Universität Stuttgart, Stuttgart, 2006.

[11] G. Möhler and A. Conkie, "Parametric modeling of intonation using vector quantization," in *Proceedings of the Third International Workshop on Speech Synthesis (Jenolan Caves, Australia)*, 1998, pp. 311–316.

[12] R. Sproat, Ed., *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach.* Dordrecht: Kluwer, 1998.