Set-up of a Unit-Selection Synthesis with a Prominent Voice

Stefan Breuer[1], Sven Bergmann[2], Ralf Dragon[3] and Sebastian Möller[4,2]
[1]IfK, University of Bonn, Germany; [2]IKA, Ruhr-University Bochum, Germany;
[3]TNT, University of Hannover, Germany; [4]Deutsche Telekom Laboratories, Berlin University of Technology, Germany
E-mail: breuer@ikp.uni-bonn.de, sven.bergmann-2@rub.de,
dragon@tnt.uni-hannover.de, sebastian.moeller@telekom.de

Abstract

In this paper, we describe the set-up process and an initial evaluation of a unit-selection speech synthesizer. The synthesizer is specific in that it is intended to speak with a prominent voice. As a consequence, only very limited resources were available for setting up the unit database. These resources have been extracted from an audio book, segmented with the help of an HMM-based wrapper, and then used with the non-uniform unit-selection approach implemented in the Bonn Open Synthesis System (BOSS). In order to adapt the database to the BOSS implementation, the label files were amended by phrase boundaries, converted to XML, amended by prosodic and spectral information, and then further converted to a MySQL relational database structure. The BOSS system selects units on the basis of this information, adding individual unit costs to the concatenation costs given by MFCC and F0 distances. The paper discusses the problems which occurred during the database set-up, the invested effort, as well as the quality level which can be reached by this approach.

1. Introduction

Unit-selection speech synthesis has become increasingly popular due to its enhanced prosodic quality and naturalness when compared to parametric or diphone synthesizers. The principle is based on the concatenation of naturally-produced speech units of variable length, avoiding signal manipulation at the concatenation points as far as possible. A sufficiently high quality can be reached when several instances of segments with different intonation are contained in the unit database; in that case, optimum units can be selected by minimizing intrinsic unit costs as well as concatenation costs.

The quality of speech synthesized with the unit-selection approach strongly depends on the availability of adequate units in the database. Thus, quality may be particularly critical when resources for setting up the database are sparse, e.g. when using resources which have not been designed for generating synthesis units, like general audio or film documents, audio books, interviews, etc. Such data material often shows a low audio quality and is not representative for the speech samples to be synthesized later; consequently, it is sub-optimal for building a unit database. Still, such resources may be the unique material when setting up synthesizers for historic and/or prominent voices, as long as voice conversion algorithms provide only limited adaptation possibilities.

In this paper, we describe the set-up process of a unit-selection synthesis for a prominent German voice, on the basis of an audio book as the only resource. As the investigation was part of a student project at Ruhr-University Bochum, the effort for cutting, segmenting and labelling the database and integrating it into our TTS system was particularly restricted. In turn, the effort for setting up such a database as well as the quality level which can be reached with such restricted resources may be estimated.

For the set-up of the database, we used a German autobiographical audio book written and read aloud by Marcel Reich-Ranicki, a well-known literature critic. This audio material was segmented with the help of an HMM-based wrapper, amended by prosodic and spectral information, and converted into a MySQL relational database. The database was used as an input to the non-uniform unit-selection approach implemented in the Bonn Open Synthesis System (Breuer et al., 2005).

Using this approach, a number of exemplary in-domain and out-of-domain speech samples were generated. The overall quality of both naturally-produced and synthesized speech samples was assessed in an informal listening-only test. The obtained results provide an indication of the quality level which can be reached by the presented approach.

2. Audio Material

The audio material was extracted from the autobiographical audio book “Mein Leben” by Marcel Reich-Ranicki, consisting of 2 CDs with extracts of the corresponding book read aloud by the author (Reich-Ranicki, 1999). Recordings were apparently made in a studio environment and available to us in standard CD format (44.1 kHz sampling frequency, 16 bit quantization). The speech is marked by a strong intonation and a prominent Polish accent. Reich-Ranicki also has a notable dental sigmatism. Nevertheless, his reading aloud sounds professional and exhibits a comparatively low level of phonetic reduction.

In order to create the database, the audio data was digitally ripped from CD and cut into individual sentences using CoolEdit Pro. The punctuation of the written version was used as a reference for the manual mark-up of sentence boundaries, apart from passages where phrasing suggested a different segmentation. Using the batch save feature of the software, 929 signal files were created based on the marked boundaries. The corresponding text was scanned from a printed copy of the book and converted into ASCII using OCR software. A spell checker was used to correct most of the recognition errors. Numbers and abbreviations were normalized manually. The resulting text was automatically split into sentences using MS Excel. The differences in text and signal segmentations were, again, resolved manually in the text version.

Lexical items were automatically extracted from this text and transcribed using the Bonn Machine-Readable Pronunciation Dictionary[1] (BOMP) for German and the grapheme-to-phoneme faculties of the diphone-based TTS system HADIFIX[1] (for unknown words).

To obtain a phonetic segmentation of the corpus, a Hidden Markov Toolkit [2] wrapper, lam4htk [3], written at the University of Hannover was used. This wrapper uses the UNIX "make" program to run the HTK tools and generate all necessary dependencies.

The lam4htk wrapper assumes one HMM with three inner states for each phone without any backward-transitions. The parameters for every state, which are transition probabilities to other states and probability densities of speech vectors, are to be determined. Therefore, sentence HMMs are built by concatenating the phone HMMs according to the transcription of the corresponding sentence. Given the recording as speech vectors, the state's parameters are chosen to maximize the joint probability when moving over all possible transitions from the first to the last state of the sentence HMM. This task is done iteratively by HTK, facilitated by the fact that there are only few possible transitions. To obtain optimal parameters over all sentences, the grammar for HTK was chosen in a way that all sentences may be recognized when given any recorded sentence. Additionally, if there are multiple possible pronunciations, HTK determines the correct one. At the end of this training step, the parameters of every phone's HMM are known. Finally, a recognition step follows which, like the training, determines a maximum likelihood and outputs the beginning and ending of the phones.

3. Database Annotation

The resulting database was used in a non-uniform unit-selection approach implemented with the BOSS open software package. In order to adapt the database to the format required by BOSS, a simple conversion from the HTK label format into BLF (BOSS label format) was necessary. To alleviate the influence of incorrectly placed segment boundaries on synthesis quality, we reduced the number of concatenation points by joining certain segments into multi-phone units according to the "phoxsy" unit specification (Breuer & Abresch, 2004). No specific intonation model was used to annotate the corpus. To provide at least a minimum of prosodic information, we manually added phrase boundaries to the label files, using the BLF specification to create a basic distinction into final-rising, final-falling or progredient-rising phrase tones.

In BOSS, all annotation beyond the level of segmental and phrasing information is stored in a specialized XML format. Tools to convert the label files into XML and augment this representation with further information are provided with the BOSS system. This includes e.g. F0 and mel frequency cepstrum coefficients (MFCC) which are used to determine transition costs in the unit-selection process. To conclude the adaptation of the corpus to the synthesizer, the XML files were automatically converted into a MySQL relational database structure which BOSS uses to access annotations at runtime.

4. BOSS Unit-Selection Synthesizer

BOSS is a non-uniform unit-selection synthesizer in the sense that utterances are concatenated from units of different linguistically or acoustically defined levels. For the Marcel corpus, we used the word, syllable and (multi-)phone as synthesis units. In the BOSS unit-selection approach, a configurable pre-selection mechanism selects a number of available candidate units on the basis of symbolic information delivered by the preceding synthesis modules. Larger units are preferred by the pre-selection process. The final sequence of units is selected by adding the individual unit costs to the concatenation or transition costs given by the MFCC and F0 distance for each potential join and then choosing the least costly path. For the purposes of this research, the synthesizer had to rely only on the information given by the text pre-processor and transcription module, including syllable boundaries, lexical stress and the phrasing indicated by punctuation to select candidate units and assign unit costs for a given input text. A CART-based duration model was employed, but was not trained on the Reich-Ranicki corpus. Instead, the data for our female voice were used. Apart from the described mark-up and prediction of phrase breaks, no intonational model was used to generate acoustic or symbolic information that would aid in the selection of synthesis units. As a consequence of this limited prosodic information, no manipulation of either duration or F0 was applied to the signal.

5. Evaluation Experiment

Although the prosodic modelling of this approach can be considered less than adequate, the results are astoundingly good. Much of the erroneous output produced can be attributed to either temporal or qualitative faults in the segmentation, that is, either a mismatch between transcription and actual realization or wrongly placed segment boundaries. The highly expressive intonation of the source material combined with the lack of modelling does not seem to lead to a higher percentage of inadequate prosodic realizations, but instead contributes to the naturalness of the synthetic voice.

In order to obtain quantitative data on the quality level which can be reached with the presented approach, an informal evaluation experiment has been conducted at Deutsche Telekom Laboratories. The experiment is informal in that only a very limited amount of speech data

[1] http://www.ikp.uni-bonn.de/dt/forsch/phonetik/hadifix/Hadifix.en.html

[2] http://htk.eng.cam.ac.uk/

[3] http://www.tnt.uni-hannover.de/

could be generated, in that the test paradigm was not tailored to a specific application scenario, and in that the test participants were selected at random in a university environment. Despite of these deficiencies, we think that the results can be used as an initial guess of the obtained quality.

The speech material used for the experiment consists of three types of utterances: Firstly, 15 sentences were extracted from the original audio data (naturally-produced speech, 'n'). Secondly, these 15 sentences were re-synthesized by excluding the respective sentences from the original inventory; thus, each sentence was generated with an inventory constructed from only 928 remaining utterances instead of the full set of 929. Like the naturally-produced ones, these samples reflect only in-domain sentences, which may be important for speech synthesized from very limited resources ('s_in'). Finally, 11 sentences were generated for a completely different domain, namely for a speech interface to domestic devices (sentences used in the evaluation of the INSPIRE system, see Möller et al., 2005). For generating these stimuli, the full inventory based on 929 sentences was used ('s_out').

The 41 stimuli were judged with respect to their overall quality in a listening-only paradigm. Because the stimuli reflect different domains, it is difficult to find an application scenario which would be plausible for all stimuli. We opted for presenting the speech stimuli as the speech output component of a smart-home assistant which would be able to read audio books and operate different domestic devices. This application scenario was introduced to the test participants in a short written introduction, but because no specific parallel tasks were given, it seems doubtful whether the application context can be seen as a reference for the obtained quality judgments. In a formal evaluation, it would be important to provide such a reference context to the participants in order to obtain valid judgments which do not only reflect the "surface form" of the test stimuli, but their communicative function as well.

Stimuli were presented to the test participants in a quiet office room via headphones (Stax Lambda Pro, diffuse-field equalization). The test was administered from a graphical interface which allowed the test subjects to play the stimuli and to provide their rating on a continuous scale through a slider. The scale was labelled with 5 describing attributes corresponding to the ACR labels for overall quality, as given in ITU-T Rec. P.800 (1996). However, because of the well-known problems of the ACR scale, the scale design was modified as described in Möller (2005). Test participants were given a short introduction on the purpose of the experiment, on the application context, and on the rating procedure. Before the actual test session, three stimuli ('n', 's_in' and 's_out') were presented to anchor the quality space. All subjective ratings were aggregated in an SPSS table for further analysis.

13 listeners (6 males, 7 females) took part in the experiment. They were between 19 and 38 years old, with a mean of 26.5 years. As the participants were recruited in a university environment, the average age was relatively low. Only 3 of them knew Marcel Reich-Ranicki and 4 of them had experience with synthesized speech, but 8 had experience with spoken dialogue systems.

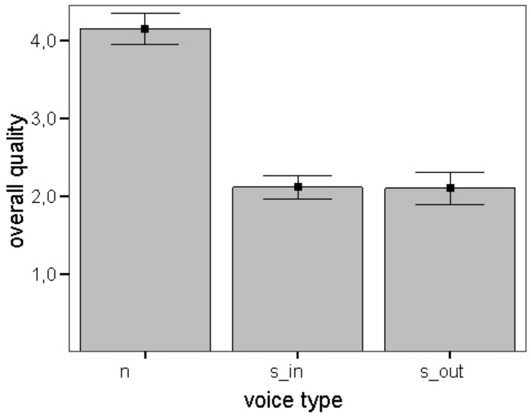

Figure 1: Mean overall quality judgments and 95% confidence intervals for naturally-produced and synthesized stimuli.

Figure 1 shows the mean overall quality ratings and 95% confidence intervals for the three types of voice ('n', 's_in' and 's_out'). Apparently the naturally-produced speech samples are rated considerably higher than the synthesized ones: Mean ratings and standard deviations are 4.15 (1.43) for 'n' compared to 2.11 (1.11) for 's_in' and 2.10 (1.23) for 's_out'. An analysis of variance shows that the difference is significant for 'n' vs. 's_in' and 's_out' ($F = 161.5$, $p < 0.001$), but a Tukey HSD post-hoc test shows no significant effects between 's_in' and 's_out' ($p = 0.995$). Apparently, the quality level of our synthesized speech is the same for in-domain and out-of-domain samples.

The large variances show that the quality largely differs between samples: The mean ratings for each sample vary between 3.51 and 4.86 for 'n', between 1.35 and 3.15 for 's_in', and between 1.46 and 2.70 for 's_out'. Interestingly, the spread is about the same for the naturally-produced and for the synthesized samples. This finding is in contrast to observations made earlier, namely that the quality of unit-selection synthesis is more prone to variances between samples than naturally-produced speech (Möller, 2005). No significant differences were found between the ratings of the participants knowing Marcel Reich-Ranicki from the media and those who did not know him.

The results show that the quality level which can be reached by our unit-selection approach is significantly lower than the one of naturally-produced speech. Considering the limited size of the unit inventory and the limited effort which could be devoted for constructing it, this finding is not surprising to us. On the other hand, it is encouraging to see that approximately the same quality level can be obtained for in-domain and out-of-domain sentences. Apparently, the size of the database is already big enough to allow for such a cross-domain extrapolation. The results should however be interpreted with care, as the set-up and the size of this initial experiment was still quite limited.

6. Conclusions

We presented a unit-selection approach to synthesize speech with a prominent voice, requiring only very limited input resources. Audio samples have been extracted from an audio book, phonetically segmented with the help of an HMM-based wrapper, and then amended by prosodic and spectral information to form a unit inventory for the BOSS unit-selection synthesizer.

The effort required was humble as compared to the effort that would have had to be put in the design, recording and labelling of a dedicated synthesis corpus. Importing and cutting the audio data took approximately four hours and the scanning, correction and splitting of the text data required about 11 hours. The automatic transcription, segmentation and adaptation to the BOSS format were performed in the course of a few days.

The main problem with the set-up of this specific unit database was perhaps the lack of control over the pronunciation of the speaker and the resulting discrepancies between automatically placed labels and actual realisations on the one hand and the transcriptions predicted by BOSS and the database on the other.

Using the described approach, in-domain and out-of-domain speech samples were generated and evaluated in an informal listening-only test. The results show that the quality level which can be reached by the presented approach is clearly inferior to the one of naturally-produced speech. On the other hand, the quality level is independent of the target domain. Thus, such a limited database seems to be sufficient to synthesize speech for different domains and applications.

The initial experiments presented here were limited by the available resources. In fact, they were conducted in the framework of a student project at Ruhr-University Bochum, and the evaluation was carried out by inexperienced supervisors (students in computer science at Berlin University of Technology). Thus, we foresee some potential for improvement even with a moderate amount of further effort. For example, the quality of the concatenated speech could be greatly improved by manually fixing at least the most severe segmentation errors (e.g. by comparing phone durations), training new duration models specific to the Marcel corpus and bringing transcriptions in line with actual pronunciations. In addition, the evaluation experiment should be repeated using more test participants and speech stimuli, and involving other target domains. Particular focus should be set on the evaluation of prosodic adequacy to find out whether audio books, with their usually rather expressive intonation are generally suitable for speech synthesis purposes

In order to illustrate the obtained quality level, a web interface has been set up which allows own synthesis examples to be generated online: http://www.ikp.uni- bonn.de/boss/bochum/.

7. Acknowledgements

The present work was performed in collaboration between the Institute of Communication Acoustics (IKA), Ruhr-University Bochum, the IfK (form. IKP), University of Bonn, the University of Hannover, and Deutsche Telekom Laboratories, Berlin University of Technology. The authors would like to thank colleagues from the respective institutions for their support during the experiments.

8. References

Breuer, S., Abresch, J. (2004). Phoxsy: Multi-phone Segments for Unit Selection Speech Synthesis. *Proceedings of the ICSLP*. Jeju.

Breuer, S., Wagner, P., Abresch, J., Bröggelwirth, J., Rohde, H., Stöber, K. (2005). *Bonn Open Synthesis System (BOSS) 3 Documentation and User Manual*. http://www.ikp.uni-bonn.de/boss/BOSS_Documentation.pdf

ITU-T Rec. P.800 (1996). *Methods for Subjective Determination of Transmission Quality*, International Telecommunication Union, Geneva.

Möller, S. (2005). *Quality of Telephone-Based Spoken Dialogue Systems*, New York NY: Springer.

Möller, S., Krebber, J., Smeele, P. (2005). Evaluating the Speech Output Component of a Smart-Home System. *Speech Communication*, 48, 1-27.

Reich-Ranicki, M. (1999). *Mein Leben*. Double CD. Munich: Sony BMG.