

MARY TTS participation in the Blizzard Challenge 2007

Marc Schröder and Anna Hunecke

DFKI GmbH
Saarbrücken, Germany
<http://mary.dfki.de>

Abstract

This paper describes the second participation of the open source MARY TTS unit selection system in a Blizzard challenge. Compared to last year's system, a number of well-defined changes have been made to the algorithm, concerning unit definition, prosody models, and signal modification. The results in this year's challenge are considerably improved, confirming that the changes were worthwhile. The paper also reports on an approach to the selection of a subset of the utterances provided, in order to build a voice with good coverage not larger than the pre-defined "Arctic" subset. Results show that this small voice is perceived slightly better than the voice we built from the Arctic subset.

1. Introduction: Motivation and Frame

The work reported in this paper is a puzzle piece in a long-term research effort on parameterisable, expressive high-quality speech synthesis. In this effort we pursue several strands, including HMM-based speech synthesis [1], which has very promising properties of reliability and parametric control, and unit selection synthesis, in which we aim to combine expressivity-based selection with signal modification [2]. The basis for research on expressivity in unit selection is a highly controllable unit selection system with a well-defined quality. The work reported in this paper provides such a baseline system, and prepares for the control of expressivity.

The benefit of the Blizzard challenge in general is the fact that it allows the research community to compare different data-driven speech synthesis algorithms on the same data. Through centrally organised listening tests, the perceptual effects of various design choices are made evident. To the extent that the differences between systems are well documented, that allows for conclusions about the effectiveness of the different methods, and thus a learning experience for all participants.

The Blizzard challenge 2007, which uses material from the same speaker as the Blizzard challenge 2006, additionally allows for *some* comparison within systems. Any major changes in perception scores between the two years are likely to be attributable to a considerable extent to changes made in the system.

For DFKI, the Blizzard challenge 2006 was our first participation in a unit selection competition [3], with moderate success: the overall impression of the listeners for our large voice, measured by Mean Opinion Scores (MOS), was somewhat below average (2.5, the average being 3.0); the intelligibility, measured by Word Error Rate (WER), was slightly worse than average (25% error compared to the average of 22% error, for native English speaking undergraduate listeners). At the workshop, we received feedback pointing out possible improvements

to the system. Some of these were implemented in time for this year's participation.

The algorithms presented in the present paper differ from the previous system in the following key points:

- diphone units, rather than phone units, are used, with a fallback to halfphones when the diphone coverage is insufficient;
- statistical prosody models are used to predict target values for F_0 and duration, and these values are used in target costs;
- signal post-processing, e.g. for F_0 smoothing, is avoided altogether.

While there have been many more changes, notably a much more efficient implementation, we consider these to be the major differences to the previous system, so that they can be expected to be key factors in explaining any differences between the results of last year's and this year's system.

The remainder of the paper is structured as follows. We first give an overview of the MARY unit selection system as it stands today, including the open source platform in general, the unit selection algorithm as such, and the voice building toolkit. We then describe the creation of voices for the Blizzard challenge, with some detail describing the selection of utterances for the small voice (voice C). We present and discuss the results before concluding on some ideas for future work.

2. The MARY unit selection system

2.1. The open source MARY TTS platform

MARY (Modular Architecture for Research on speech sYnthesis) is a platform for research, development and teaching on text-to-speech synthesis. Originally developed for German, it was extended to US English by incorporating some TTS modules from the FreeTTS project [4], and, as the result of a student project, to Tibetan. MARY uses an XML-based representation format for its data, which makes it possible to access intermediate processing states, and to connect it to other XML-based processing components.

Apart from being a research platform, MARY is also a stable Java server capable of multi-threaded handling of multiple client requests in parallel.

The design is highly modular. A set of configuration files, read at system startup, define the processing components to use. For example, the file `german.config` defines the German processing modules, `english.config` defines the English modules, etc. If both files are present in the configuration directory, both subsystems are loaded when starting the server. Each synthesis voice is defined by a configuration file: `german-mbrola-de7.config` loads the MBROLA

voice de7, english-arctic-jmk.config the unit selection voice built from the Arctic recordings of speaker jmk [5], etc.

Each synthesis module has an input and an output format, which can be flexibly defined. This makes it extremely easy to define pipeline architectures for processing any given input format into one or more output formats, without explicitly stating the required chain of modules. Starting from the input format specified for the system input (e.g., plain text, SSML [6], etc.), the TTS system searches a path through the available processing components until it arrives at the requested output format (e.g., audio). Although this is a very simple mechanism for specifying a component architecture, it seems to be sufficient for the processing requirements of a TTS system.

For the generation of audio, MARY includes the concept of a collection of waveform synthesisers; these are defined in an extensible way through the MARY configuration files. Currently, the list of available waveform synthesisers includes the MBROLA diphone synthesiser; an LPC-based diphone synthesiser provided by FreeTTS; the MARY unit selection synthesiser covered in the present paper; and an experimental interpolating synthesiser, creating intermediate voices from two existing unit selection voices [2] using a spectral interpolation algorithm [7].

The architecture of the MARY platform as well as the English and Tibetan processing components are available under a liberal BSD-style license. The German processing components are available free of charge under a research license. By permission from the MBROLA team, MBROLA binaries and voices are provided with MARY under the MBROLA license.

The system runs under Windows, Linux, Solaris, and Mac OS X. A comfortable graphical installer can be downloaded from the MARY website. During installation, users can indicate which components they want to install; only these components are downloaded from the MARY page.

In order to avoid misconfigurations, the configuration files define a number of dependencies, which are checked automatically at every system startup. If a component is found to be missing, the system offers to download it from the MARY website.

2.2. Unit selection in MARY

The unit selection system in MARY implements a generic unit selection algorithm, combining the usual steps of tree-based pre-selection of candidate units, a dynamic programming phase combining weighted join costs and target costs, and a concatenation phase joining the selected units into an output audio stream.

Units to concatenate are uniform. The Blizzard 2006 system [3] used phoneme units. After getting feedback at the Blizzard Challenge Workshop 2006, we switched to diphone units, because joining in the mid-section of phonemes is expected to introduce less discontinuities than joining at phoneme boundaries. For each target diphone, a set of candidate units is selected by separately retrieving candidates for each halfphone through a decision tree, and retaining only those that are part of the required diphone. When no suitable diphone can be found, the system falls back to halfphone units.

The most suitable candidate chain is obtained through dynamic programming, minimising a weighted sum of target costs and join costs. Both are themselves a weighted sum of component costs. Target costs cover the linguistic properties of units, and the way they match the linguistically defined target. A sec-

ond major change compared to the Blizzard 2006 entry is the use of acoustic target costs. These are currently used for comparing a unit's duration and F0 to the ones predicted for the target utterance by means of regression trees trained on the voice data. In the future, we intend to use acoustic target costs to also cover expressivity-related acoustic measures, such as spectral tilt or other robust measures of voice quality.

Join costs are computed as a weighted sum of F0 difference and of spectral distance, computed as the absolute distance in 12-dimensional MFCC space. We had experimented with a step function for the F0 penalty, based on the reasoning that small F0 deviations can be corrected by a smoothing algorithm [3]; currently, we are using a linear cost function instead and avoid signal post-processing as it seems to degrade the overall quality.

Like all unit selection systems, we face the challenge of determining appropriate weights for the individual target and join cost components. As we have not yet developed a principled way of determining these weights, we have set a number of ad hoc values through iterative listening and adapting. The resulting weights give slightly higher importance to join costs than to target costs, a higher importance to F0 continuity than to spectral continuity, and a higher importance to duration and F0 targets than to phonetic context.

After the chain of units minimising these costs is determined, the units are retrieved from a timeline file and concatenated using overlap-add of one pitch period at the unit boundaries. The timeline file currently contains uncompressed PCM audio data, but is designed in a way that makes it easy to use more efficient encodings in the future.

The system is reasonably efficient: it synthesises speech about ten times faster than real-time on a recent Core 2 Duo processor. Decision trees and feature vectors required for the cost computation are held in memory; audio data is retrieved from a file after selection.

2.3. The voice creation toolkit in MARY

We are in the process of developing a toolkit for creating voices for MARY. We originally used the Festvox tools [8], and we continue to be deeply grateful to their creators for making them available to the community. However, it appears that some aspects of Festvox are tightly linked to the Festival system, and we felt that in the long run, the gain in control and flexibility justifies the development of our own voice creation toolkit.

The system combines an extensible list of "voice import components" in a graphical interface which is currently still very simple (see Figure 1). The user can select a series of import components, which are run in sequence. A progress bar is shown for the component which is currently running. After successful completion, the component is coloured in green; if processing fails, it is displayed in red, and processing of subsequent components is aborted. Configuration of non-default file system paths and special settings for the components is done via command-line options.

The voice import components that are currently available include components for automatic labelling using Sphinxtrain [9]; for importing text files in Festvox format; for predicting unit features with MARY; for making sure the unit labels and the feature chain predicted by MARY are properly aligned; for pitchmarking using Praat [10]; for the conversion of data into the compact format required by the MARY unit selection runtime system; for building classification trees for candidates using the wagon tool from the Edinburgh speech tools [11]; for pruning outliers from the generated trees; and for creating re-

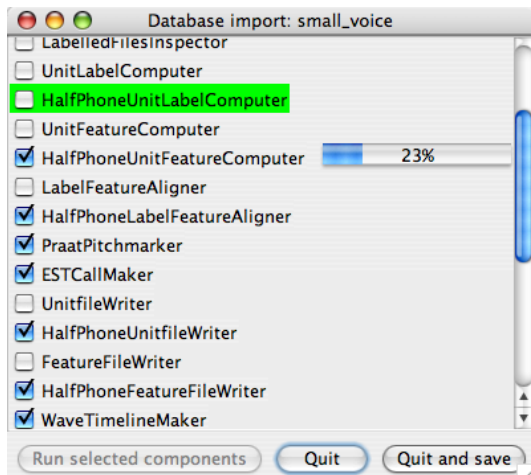


Figure 1: The MARY voice creation toolkit at work. In the situation shown, half-phone unit labels have been created successfully, unit features are being computed, and a number of components are scheduled for subsequent execution.

gression trees for duration and F0.

One of the most time-consuming tasks is the training of classification trees for the prediction of candidate units. Similarly to [12], we use acoustic distance between units as the impurity measure, and run wagon based on distance tables. In order to speed up the process on a multi-processor machine, the MARY CartBuilder component can run several wagon processes in parallel. Given the fact that the computation of acoustic distances is currently done in a single Java process, there is a limit to the number of wagon processes that should reasonably be started in parallel; we have experienced considerable speedup with running 3-5 wagon processes alongside one Java process on an 8-processor machine.

The MARY voice creation toolkit currently requires a considerable amount of expert knowledge in order to set paths correctly via command-line options and to select the right components for the task at hand. We intend to develop a more intuitive system providing groupings of the components that are usually required for a given task. For example, components working with halfphones are required for creating the necessary files to build classification trees for pre-selection of candidate units, but phone-sized units are needed for training regression trees for the prediction of duration and F0.

3. The Blizzard voices

As in previous Blizzard challenges, participants’ task consisted of building synthesis voices from recordings provided centrally. This year’s database consisted of 6579 sentences (corresponding to 477 minutes of speech data, i.e. just under eight hours) kindly provided by ATR. The speaker is the same as in last year’s database, but the selection of sentences is different. The database includes a subset of 1032 sentences (49 minutes, approximately 1/10th of the total) called the “Arctic” subset. Three voices were to be built: one from the full set of recordings (voice A); one from the Arctic subset (voice B); and one from a self-chosen subset of sentences (voice C). The subset used for voice C could be chosen from all sentences. Full sentences needed to be chosen; the total duration of speech in the chosen subset was limited to the total duration of speech in the Arctic set.

3.1. Preselection of sentences and transcription of unknown words

One of the lessons of the Blizzard Challenge 2006 for us was the importance of pre-selecting the material, to make sure that automatically predicted pronunciation matches the spoken utterances.

We used the MARY system to identify unknown words; these were either added manually to the pronunciation lexicon, or the respective sentences were discarded.

1030 words were unknown to the MARY pronunciation lexicon. These included 590 foreign words (mostly Japanese and Spanish), stemming from 521 sentences, which were discarded from the set. The remaining 540 unknown words and their transcriptions generated by the letter-to-sound rules were inspected by a trained phonetician; for 151 of them, the transcription was manually corrected and added to the pronunciation lexicon, including some abbreviations which had not been expanded by the MARY text normalisation component; the remaining unknown words were judged to be appropriately transcribed by the letter-to-sound rules.

A custom sub-corpus was selected from the full set as the data for voice C (see Section 4 for details).

3.2. Applying the MARY voice preparation tools

The MARY voice preparation tools described above were applied to the respective subsets of utterances for the creation of the voices A, B and C.

The data was automatically labelled with the phonemes predicted from text by the MARY phonemisation component after transcriptions for unknown words had been added. Sphinxtrain was used to force-align this phoneme chain with the recording files. No verification of the labelling quality was done at this stage, but some outliers were removed in the pruning stage after building pre-selection trees. This is a potential area of improvement (see Conclusion below).

Pitch marks and pitch-synchronous MFCC vectors were computed with Praat and the Edinburgh Speech Tools (EST), respectively. Linguistic feature vectors were predicted with the MARY system; acoustic unit features (F_0 and duration) were added based on pitchmarks and on automatically labelled phoneme boundaries. Join cost features (F_0 and MFCCs) were measured at the first and last frame of each halfphone unit.

Pre-selection trees for halfphone units were created in a two-step procedure. We first specified a “top-level” tree by hand, which organises all units into “top-level leaves” according to phonetic properties such as phoneme identity, stress status, voicing etc. This resulted in a top-level tree with about 1000 leaves of up to 3400 units each. Larger top-level leaves were avoided because we observed an extreme increase in processing time needed for automatically growing classification trees from larger unit sets.

Second, we applied the EST tool *wagon* in an automatic tree-growing procedure for each of the top-level leaves, using acoustic distance between units as the impurity measure. The acoustic distance between two units was measured as the weighted sum of differences in duration, F0, and spectral difference computed as the Mahalanobis distance of MFCC vectors. The mapping of frames from the two units for computing the spectral distance was performed using linear time scaling.

Regression models for F_0 and duration were computed from phone models, using *wagon* with default settings from the Festvox documentation.

3.3. Pruning

As one measure to reduce the effect of wrong labelling, we have introduced a first pruning algorithm into the voice building process, looking for outliers in the classification tree used for pre-selecting candidate units. Outliers were identified for each leaf based on a number of criteria, including the likelihood of each unit computed by wagon as well as the energy. For example, units labelled as silence but with a high energy are considered outliers and removed.

Due to the preliminary stage of this processing component, conservative settings were chosen. Only the pre-selection tree of the full voice (voice A) was pruned; in total, 1% of the units were removed.

3.4. Tuning of weights

Despite early publications reporting on attempts to determine the weights for target and join costs based on objective criteria [13], it seems to be common practice to tune weights by hand. This seems to be due to the difficulty to find acoustic measures that reflect perceptual impression. While we intend to investigate the question in the future, in the current system we also tuned weights manually. In a trial-and-error procedure, weights were tuned such that acoustic target costs have a slightly higher weight than linguistic target costs, and join costs have a slightly higher weight than total target costs.

4. Selection of a custom sub-corpus

The corpus of Voice C was selected with a greedy algorithm. That means, at each step the sentence that gets the highest score according to some criteria is selected to be added to the cover set. The algorithm stops when the duration of all sentences in the cover reach the maximum duration of 2914 seconds (the total speech duration of the Arctic set). The major parameters of the algorithm are the coverage definition and the sentence score.

4.1. Coverage definition

The definition of coverage determines which units are wanted in the final set. It depends on the definition of the units.

Units are represented as vectors consisting of four features. For each phone, there is one feature vector. The four features are phonetic identity, phonetic identity of the next phone, phone class of the next phone and prosodic characteristics of the current phone.

The English phoneset that was used defines 42 different phones. Thus, there are $42 * 42 = 1764$ different possible diphones.

The concept of phone classes was introduced to reduce the number of possible diphones. The idea behind this is that the transitions in the middle of two diphones are similar if the second parts of the diphones are similar phones. For example, the transitions from a vowel to alveolar consonants will be similar, no matter which alveolar consonant it is. But they will be distinct from the transitions of that vowel to a velar consonant. For the consonants, the place of articulation is more important for the transitions than the manner. The same is true for the vowels: for example, [i] and [y] have the same place of articulation and hence will have the same transitions leading to them. In this manner, 21 phone classes were defined, which thus reduces the number of possible diphones from 1764 to $42 * 21 = 882$.

For the prosodic characteristics of a phone, six different

	full	Arctic	voice C
Number of sentences	5879	1028	836
simple diphones	81.65%	77.12%	76.66%
simple diphones & prosody	53.50%	34.71%	44.66%
clustered diphones	86.30%	81.77%	86.30%
clustered diphones & prosody	60.65%	41.54%	59.39%

Table 1: Distribution statistics of the three corpora

prosodic types were defined: unstressed, stressed, pre-nuclear accent, nuclear accent, phrase final high and phrase final low. The accents and phrase final tones were computed on the basis of ToBI predictions from text.

With this we have two different definitions of coverage:

- **simple diphones:** the cover consists of all combinations of phone, next phone and prosodic type.
- **clustered diphones:** the cover consists of all combinations of phone, phone class of next phone and prosodic type.

4.2. Sentence score

For each unit, a score determines how “useful” the unit is for the selected set. For each sentence, the score is the normalized sum of the scores of the units.

The score of a unit is basically the product of two different weights: frequency weight and “wanted” weight. The frequency weight reflects the frequency of the unit in the corpus. The “wanted” weight reflects how much a unit is “wanted” in the cover: If there is already an instance of this unit in the cover, the wanted weight will be lower than if there is no instance in the cover. The wanted weight can have a different setting on the three levels phone, next phone/next phone class and prosody. This way the wanted weight also determines what is more useful: new phones, new diphones or new prosodic types.

For the frequency weight, three settings are considered: 1 (which means no consideration of frequency), relative frequency (which gives a preference for the more common units) or the inverse of the relative frequency (which gives a preference to the rarest units).

An additional dimension is added by the setting for the decrease of the wanted weight: Each time a unit is selected for the cover set, the wanted weight for this unit is divided by a certain number, to reflect the fact that we already have this unit and do not necessarily want another instance of it. The higher this number is, the less useful new units that are already in the cover will be.

4.3. The Voice

The algorithm was run several times with different settings. The most important setting variation was the definition of coverage: simple or clustered diphones. Another variation was the setting of the frequency weight to the three possible settings. The setting of the wanted weight was varied between 100, 10 and 1, and the number by which the weight is divided was varied between 10000, 1000, 100 and 10. Additionally, there were tests with restricting the sentence length.

The results indicated that, in general, using the simple diphone coverage definition maximizes both simple and clustered coverage, but using the clustered diphone coverage definition only maximizes clustered coverage. Also, the use of the inverse frequency generally led to better results than the normal

frequency. Generally, the restriction of sentence length did not lead to good results in the test.

The selected sentences with the best distributions were chosen to build three different voices. Of these, the best voice was selected on the basis of informal listening tests. For this final voice the settings were: clustered coverage definition, inverse frequency, wanted weight of 1 on all levels, divided by 100, and no restriction on the length of the sentences.

Table 1 shows the diphone distribution of the corpus of the final voice (voice C) in comparison with the distributions of the corpora of the other two voices that were submitted. It can be seen that, on the one hand, voice C has fewer sentences than the Arctic voice and also a slightly lower simple diphone coverage. But on the other hand, the percentage of prosodic variations of both simple and clustered diphones is higher for voice C than for the Arctic voice, and clustered diphone coverage is as high as for the full corpus.

5. Results

Our goal in this year’s challenge was to make progress towards being perceived as good as the average of all systems, given that we were a bit below average (at 2.5 MOS and 25% WER) last year. Our goal has been surpassed – our system performed better than average on both MOS and WER with all three voices (see Figs. 2, 3, 4).

The box plot in Figure 2 shows the median and quartiles of MOS ratings by all listeners for voice A across systems. Systems are ordered on the X axis according to their mean MOS for voice A; the DFKI entry is system C.

Figure 3 shows a simplified view of the same data for all three voices, comparing the DFKI entry with the average of all participating systems. It can be seen that all three voices are rated better than the average of all participating systems. It is also visible that our voice C has a better rating than our voice B.

Figure 4 shows Word Error Rate (WER), for native speakers of English only. Figures for non-native listeners were very high (close to 50% for most systems), indicating that the task may have been too difficult for non-native listeners to provide informative differences between systems. For this reason, we consider only WER for native English listeners. It can be seen that the intelligibility of the DFKI voices is considerably better than the average of all systems; the WER for our full voice is only half as high as for our two small voices.

6. Discussion

Given the similarity in procedures and data, it seems justifiable to compare the MOS and WER figures from the Blizzard Challenge 2006 and 2007, at least informally. Such a comparison can lead to qualitative rather than quantitative insights: concretely, we would like to know if our system has become better by the modifications we made to it since last year.

The average MOS across all systems is nearly unchanged (2.9 for the full voice, both this year and last year). The rating of our full voice (3.2) is considerably higher than last year (2.5), indicating that the changes we made to the system led to a substantial improvement of perceived naturalness.

Figures for WER for native speakers this year may best be compared to the “undergraduate” figures of the Blizzard Challenge 2006, because the undergraduates were native speakers of English. This comparison shows nearly no change for the average of all systems (WER for the full voice: 22% in 2006,

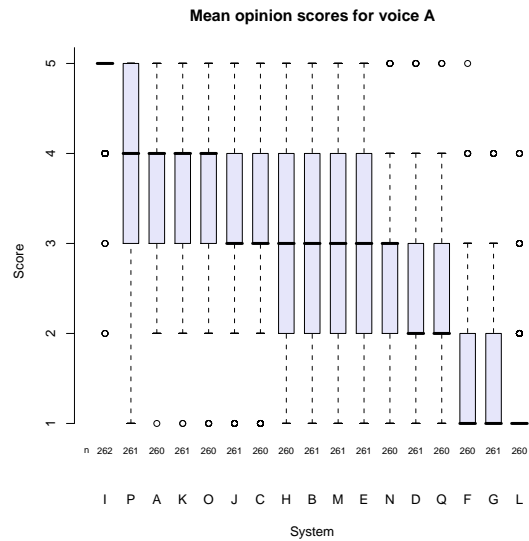


Figure 2: Mean Opinion Scores (MOS) of all systems, for the full voice (voice A). The DFKI entry is system C.

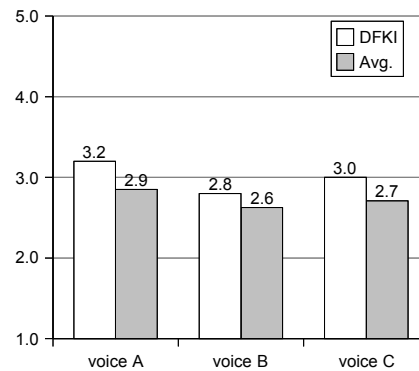


Figure 3: Mean MOS for the three DFKI voices, compared to the average of all participating systems. Higher scores are better.

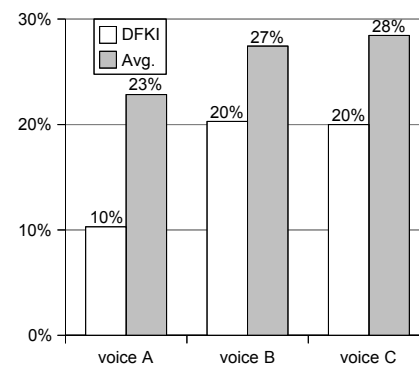


Figure 4: Mean Word Error Rate (WER) of native English listeners, for the three DFKI voices, compared to the average of all participating systems. Lower scores are better.

23% in 2007). However, the WER for the DFKI system marks a sharp decrease (from 25% in 2006 to 10% in 2007, for the full voice). Again, these figures may not be fully comparable in a quantitative way, due to the fact that voices were built from a similar but not identical data set, and that the basis for assessing WER was not the same (modified rhyme test and semantically unpredictable sentences in 2006, only semantically unpredictable sentences in 2007). Nevertheless, the qualitative observation that the WER is markedly lower in 2007 confirms that our system modifications improved intelligibility.

These observations seem to indicate that the modifications made (using diphone units and acoustic target costs, and avoiding pitch smoothing) have increased the naturalness and the intelligibility of the DFKI system considerably.

Comparing voices A, B and C, it can be seen that the use of only about 10% of the speech data (477 minutes for voice A, 49 minutes for voices B and C) had a large effect on the intelligibility of our voice, visible in a doubling of the WER (Figure 4). The naturalness rating is also affected, but not as drastically (Figure 3).

Comparing our voice C to the default Arctic voice B, it can be noted that the WER is the same for both voices, but that the MOS score is a bit better for voice C. This may be the result of our selection strategy, aiming for prosodic richness rather than the best possible diphone coverage. It would be interesting to compare the perceptual effects of various selection strategies, in order to verify whether any systematic correlations exist, e.g. between diphone coverage and WER or between prosodic richness and MOS.

7. Conclusion

The results for DFKI in this year's Blizzard challenge are considerably improved compared to the first participation in 2006. This difference is most probably due to the modifications made to the system between the two tests: the use of diphone units rather than phone units; the use of acoustic target costs based on regression trees for F_0 and duration; and the avoidance of pitch smoothing.

The test results were better than the average of all participating systems, for both MOS and WER for all three types of voices (A, B, and C). This confirms that the MARY system can now be considered to be a state-of-the-art unit selection system in its own right. We conclude that the MARY system is now a suitable baseline system for our research on expressive speech synthesis.

Nevertheless, many ideas for improvement remain to be explored. This includes the automatic assessment of quality – heuristics may help find likely problems in the automatic phoneme labelling, which can then either be discarded or presented to a human labeller for inspection. Prosody models should be computed in a normalised representation (e.g., z scores), to become independent from concrete Hz and ms values, which will make the models reusable for different voices. Alternatively, it may be worth considering model-based prosody prediction (e.g., [14], [15]) rather than purely statistical regression trees. Finally, various approaches to prosody modification will need to be carefully investigated – the observation that results improved when not using the one method we tried should not be generalised too quickly. We will compare various approaches, and assess their effects through listening tests. In parallel, we will continue to investigate HMM-based synthesis, which has inherent properties beneficial for expressivity control.

8. Acknowledgements

The work reported here was supported by the EU project HUMAINE (IST-507422), by the DFG project PAVOQUE, and by the ProFIT project IDEAS4Games. The authors would like to thank Sacha Krstulović for helping with the definition and implementation of file formats and voice import components; Mat Wilson for the design and implementation of a GUI for recording new voices; and Maximilian Kwapil for implementing the pruning algorithm for outliers in the classification tree.

9. References

- [1] S. Krstulović, A. Hunecke, and M. Schröder, "Investigating HMMs as a parametric model for expressive speech synthesis in German," in *Proc. ICPHS*, Saarbrücken, Germany, 2007.
- [2] M. Schröder, "Interpolating expressions in unit selection," in *Proc. 2nd International Conference on Affective Computing and Intelligent Interaction (ACII'2007)*, Lisbon, Portugal, to appear.
- [3] M. Schröder, A. Hunecke, and S. Krstulović, "OpenMary – open source unit selection as the basis for research on expressive synthesis," in *Proc. Blizzard Challenge'06*, 2006.
- [4] "Freetts 1.2," <http://freetts.sourceforge.net>, 2005.
- [5] J. Kominek and A. W. Black, "CMU ARCTIC speech databases," in *Proc. 5th ISCA Speech Synthesis Workshop*, Pittsburgh, PA, USA, 2004, pp. 223–224.
- [6] M. R. Walker and A. Hunt, *Speech Synthesis Markup Language Specification*, W3C, 2001. [Online]. Available: <http://www.w3.org/TR/speech-synthesis>
- [7] O. Turk, M. Schröder, B. Bozkurt, and L. Arslan, "Voice quality interpolation for emotional text-to-speech synthesis," in *Proc. Interspeech*, Lisbon, Portugal, 2005.
- [8] A. W. Black and K. Lenzo, "Festvox: Building synthetic voices, edition 1.6," Language Technologies Institute, Carnegie Mellon University, PA, USA, Tech. Rep., 2002. [Online]. Available: <http://www.festvox.org>
- [9] R. Mosur and K. A. Lenzo, *Sphinx-II User Guide*, CMU, <http://cmusphinx.sourceforge.net/sphinx2/doc/sphinx2.html>.
- [10] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." <http://www.praat.org>, 2007.
- [11] S. King, A. W. Black, P. Taylor, R. Caley, and R. Clark, "Edinburgh speech tools library," http://www.cstr.ed.ac.uk/projects/speech_tools, 2003.
- [12] A. W. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *Proc. Eurospeech*, Rhodes/Athens, Greece, 1997.
- [13] A. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP*, Atlanta, Georgia, 1996.
- [14] H. Mixdorff and H. Fujisaki, "The influence of focal condition, sentence mode and phrase boundary location on syllable duration and the F0 contour in German," in *Proceedings of the 14th International Conference of Phonetic Sciences*, San Francisco, USA, 1999, pp. 1537–1540.
- [15] S. Prom-on, Y. Xu, and B. Thipakorn, "Quantitative target approximation model: Simulating underlying mechanisms of tones and intonations," in *Proc. ICASSP*, Toulouse, France, 2006, pp. I-749–752.