Multilingual MARY TTS participation in the Blizzard Challenge 2009

Marc Schröder, Sathish Pammi, Oytun Türk

DFKI GmbH, Saarbrücken and Berlin, Germany
{firstname.lastname}@dfki.de

Abstract

The paper describes the Blizzard Challenge 2009 participation of MARY TTS, an open-source TTS system using a unit selection voice. We briefly outline the new language support framework we provide so that people can add support for their languages to MARY TTS, and describe how that framework was used for building a Mandarin Chinese system and voice. The system performs well for English and reasonably for Chinese.
Index Terms: speech synthesis, unit selection, multilingual, open source

1. Introduction

The present paper describes the fourth participation of the MARY TTS system in a Blizzard Challenge. After an initial phone unit selection system with moderate quality [1], we moved to diphone units which improved the quality substantially [2]. Through an oversight, the pre-selection CART trees in our 2008 system [3] discarded the majority of the neutral units for the full-size voice, giving lower naturalness Mean Opinion Scores (MOS) to our full-size voice compared to our Arctic voice. The error allowed us to conclude that the mere size of the database was indeed not the dominant factor, but that consistency and neutrality of the database speech material supported the perceived quality.

The present participation is on the one hand a gradual evolution from the previous systems. We avoid problems encountered in the past – this year, by barring unit pre-selection altogether from our synthesis workflow. Furthermore, we try out minor new things – this year, a stylisation of the F0 contour before computing the F0 values at unit boundaries, used for computing the F0 aspect of join costs. Also, we try a simple formant enhancement and gain adjustment algorithm for the "spokes" task on telephone-based speech.

A novelty lies in the multilingual support of the MARY TTS platform, which allows us to try a first participation in the Mandarin Chinese section of the Challenge. Without speaking Chinese ourselves, we attempted to build a voice from the data provided and from existing resources.

The paper is organised as follows. Section 2 describes some key aspects of the MARY TTS system, notably the new language support which will be made publicly available for adding new language components to the MARY TTS system. We also describe the characteristics of the system entered into the Blizzard Challenge 2009. Section 3 describes the results of our entry in the challenge, wich are then discussed in Section 4.

2. The MARY TTS system

The MARY TTS platform is an open-source, modular architecture for building text-to-speech systems, including unit selection and statistical parametric waveform synthesis technologies. It has been described in detail elsewhere [3][4]. The present paper describes only the aspects relevant in the current context.

2.1. New language support

The upcoming release 4.0 of MARY TTS will provide support for adding support for new languages and for building the corresponding voices. Whereas high-quality support of a language will usually require language-specific processing components, it is often possible to reach at least a basic support for a language using generic methods [5]. We aim to provide the tools and generic reusable run-time system modules so that people interested in supporting a new language for MARY TTS can do so.

The steps required to add support for a new language from scratch are illustrated in Figure 1. Two main tasks can be distinguished: building at least a basic set of natural language processing (NLP) components for the new language, carrying out tasks such as tokenisation and phonemic transcription; and the creation of a voice in the new language.

For both tasks, our workflow starts with a substantial body of UTF-8 encoded text in the target language, such as a dump of the Wikipedia in the target language. After extracting the actual text without markup, as well as the most frequent words, the first step is to build up a pronunciation lexicon. We use an XML file describing the allophones that can be used for transcription, providing for each allophone symbol the phonetic features that are to be used for characterising the phone later. Using a special transcription GUI, a language expert can transcribe as many of the most frequent words as possible using the allophone inventory. A "train and predict" button in the GUI is used to automatically train a simple letter-to-sound algorithm, based on decision trees, and predict pronunciations for the untranscribed words in the list. Furthermore, as a simplistic approximation of a part-of-speech tagger, it is possible to mark function words in the list.

With this minimal manual input for a new language, a simple NLP system can be built for a new language, using a generic tokeniser and a rule-based prediction of symbolic prosody.

Once the NLP component is running, the task of creating a voice can be pursued (right branch in Figure 1). First, a recording script providing good diphone and prosodic coverage (Becker 2006) is selected from the text collection. A "feature maker" component annotates each sentence in the database with diphone and prosody features to be used in a greedy selection. The resulting collection of sentences can be used as the recording script for voice recordings with our tool "Redstart". The recorded audio files can then be processed by our voice import tools which generate a unit selection and/or an HMM-based voice, as well as speaker-specific prediction components for acoustic parameters. If, during the voice-building process, force-aligned transcriptions were manually corrected, it is also possible to predict speaker-specific pronounciations.

The toolkit has been successfully applied to the creation of Turkish TTS at DFKI. It has also served as the basis for the support of Mandarin Chinese in the present Blizzard participation. A public beta release is scheduled for September 2009.

10.21437/Blizzard.2009-6

2.2. Changes to the English system since last year

The new language support also benefited our British English system. Rather than disguising the British English system as a US English system, as we did last year, we could this year properly distinguish between generic English components and country-specific components and data.

The changes to the previous system are well-defined.

In order to make sure we do not repeat the mistake of discarding suitable units, we completely avoided pre-selection of units in the before the dynamic programming step of unit selection. This slowed down the synthesis process substantially, especially for the large voice. With preselection, the time to synthesise a sentence is of the order of a few hundred milliseconds. Without preselection, the time needed to synthesise the 1415 English test sentences was 3459 seconds for the full voice, EH1, with a median synthesis duration of 1.9 seconds but with some outliers taking up to 45 seconds to synthesise, and 523 seconds for the Arctic voice, EH2 (median 0.3 s, outliers up to 4.0 s). It should be possible to speed up this process by introducing a beam search in the dynamic programming step.

Furthermore, after the experience that using expressive speech material decreased the naturalness (at least with neutral text material as used in this Challenge), we have decided to discard all expressive sections of the database (*carroll*; *address*; *spelling*; *emphasis*) and build a voice solely from the unexpressive parts (*unilex*; *arctic*; *all_news*).

Whereas we had observed a small advantage of using a statistically trained join model, for simplicity we reverted to the simple join cost based on absolute distances between F0 and MFCC parameters at unit boundaries this year. However, we changed the source of F0 values used: whereas previously we had used the raw output of the pitch tracker, this year we fitted third-order polynomials to the pitch contour of each syllable, and used the estimated F0 values for the F0 component of the join costs.

2.3. The Mandarin Chinese system

Mandarin is a standard form of Chinese, although there are other dialects. The corpus provided by iFlytek contains recorded speech of 6000 spoken utterances in Mandarin from the news domain, and in addition to that it has text transcription in GBK encoding, pinyin transcription with syllable tones, parts of speech information and word boundaries.

A set of generic modules, to support a new language, are provided by MARY as described in Section 2.1. We converted GBK encoded text transcription to UTF-8 encoded text as it is supported by MARY. Following [8], we defined set of allophones and their phonetic features for Mandarin. As Chinese is a tonal language, the syllable tone is used as an extra feature despite the fact that other languages in MARY do not use this feature.

The Mandarin pronunciation lexicon is prepared with the given pinyin transcription. Many words have multiple pronunciation in terms of syllable tone. Taking majority number of occurrences used as a criterion, a single pronunciation was chosen to prepare the lexicon since the current MARY framework supports only a single pronunciation per word. For example, the word 总统 is pronounced 20 times as '*zong2tong3*' among 26 instances in the training data. Other instances are '*zong2tong2*' and '*zong2tong0*'. In this case the lexicon uses 'zong2tong3' as the pronunciation. Among 18448 possible words in the training data, around 3.3% of words are having multiple pronunciations. We did not create a component to predict pronunciations for unseen words; instead, we added the pinyin transcriptions for the test sentences to the pronunciation lexicon.

We also did not use a word boundary tokeniser, but relied on the word boundaries given for both the training and the test sentences.

Our generic part of speech tagger uses the simple distinction whether or not a word is a function word. So, we mapped conjunctions, pronouns, particle and auxiliary words to function words and remaining to content words.

Once generic NLP modules are running for Mandarin, the voice building process is same as described in Section 2.2. The highest weight was given to syllable tone feature as compared to all other features while computing target cost for unit selection.

Altogether, the given 949 Mandarin test sentences are synthesized in 1367 seconds, where a median synthesis duration is 1.4 seconds but some outliers had taken up to 4 seconds.

2.4. Acoustic modifications for telephone speech

We have tried two simple ideas with an aim to enhance text-to-speech output for telephone channels. The first idea is based on the enhancement of lower frequency formants using a combination of line spectral frequency based modifications and frequency domain processing. For this purpose, line spectral pairs that are no more apart than 300 Hz and that lie in the frequency range 1 KHz and 2.5 KHz are detected using conventional linear prediction techniques. The distance between each LSF pair is then reduced by a specific amount (15% in our case) by moving the two LSFs towards the pair mean using the frequency domain processing approach reported in [6]. Since close LSF pairs are related to formants [7], this approach sharpens the formants in the 1-2.5 KHz frequency range. Our second idea was to increase the relative gains of consonants and decrease the relative gains of vowels. Full gain modification is applied only in the steady segment of each phoneme starting from the 25% of the total duration of the phoneme and ending at 75%. For the beginning part of the phoneme, the gain modification factor is gradually changed from 1.0 to the target amount. Similar gradual change from the target amount to 1.0 was realised in the end part as well. The target gain factor was 1.5 for consonants and 0.7 for vowels. The gains of plosives and silent segments were not modified. Although we were able to confirm slight increase in intelligibility through informal listening of a few sentences, the overall results showed that these simple methods resulted in lower intelligibility and naturalness.

3. Blizzard results

We participated in the tasks EH1, EH2, ES2, and MH. In the Blizzard listening test, the MARY system is identified by the letter 'I'. MOS and WER rates for all listeners are summarised in Table 1.

	EH1		**EH2**		**ES2**		**MH**	
	mos	**wer**	**mos**	**wer**	**mos**	**wer**	**mos**	**cer**
MARY	3.1	0.26	3.3	0.2	2.8	0.54	3.3	0.33
avg.	2.6	0.27	2.6	0.24	2.6	0.44	3.3	0.24

Table 1: Average ratings of naturalness (Mean Opinion Score, mos) and intelligibility (Word Error Rate, wer for English, Character Error Rate, cer for Mandarin) across all listeners, for the MARY system (system 'I') and the average of all systems (excluding the natural sample 'A').

Our English system (EH1 and EH2) is doing quite well on naturalness. According to the significance tests distributed by the Blizzard Challenge organisers, the naturalness scores for our EH1 voice are significantly worse than the natural recordings

(A) and the systems S and K, and significantly better than all other systems except for system B, for which the difference in ratings with our system is not significant. For the EH2 system, the same two participants (K and S) have higher mean naturalness ratings than our system; however, this difference is not statistically significant.

Intelligibility is close to average for EH1 and better than average for EH2. While several systems have better (lower) word-error rates than our systems, only for one system this difference is significant: System S in task EH1, and system C in task EH2.

Surprisingly, both the naturalness ratings and the intelligibility of the EH2 system are better than the larger voice EH1. Given the fact that there are no cross-voice significance tests, it is difficult to say whether these differences are significant. They certainly are unexpected: as the speech material used in EH2 is a subset of the speech material used in EH1, it should be assumed in theory that the unit selection algorithm has at least the same, and normally better, coverage in case EH1 compared to EH2.

The improvements of our full voice (naturalness 3.1) compared to last year (naturalness 2.8) can probably be attributed to the fact that we use a better selection from among the available speech database – exclusively unexpressive material this year, whereas last year we used predominantly expressive material. The improvement of the arctic voice (from 3.2 to 3.3) may either be random error or may be due to the more robust way of estimating F0 for the join costs.

For the telephone speech task, ES2, it seems that our method for enhancing intelligibility was not successful: the word-error rate of our system is worse than the average of all systems, whereas for the unfiltered version of the voice, EH1, we were very close to the average. Also, the MOS suffered somewhat, bringing us closer to the average of all systems. For some reason, that average is not worse for the task ES2 than for EH1, despite undeniably worse acoustic quality. Possibly, listeners have compensated their judgements for the effects of the telephone channel which they know well from everyday life.

For Mandarin Chinese, finally, our system reached a naturalness score at the same level as the average of all systems. However, the word-error rate was substantially worse than the average; actually, it was the highest of all systems.

4. Discussion

Our good results for English confirm the fact that our simple approach to unit selection continues to compare well in the overall landscape of participating systems. Among the simple rules that seem worth following are: selecting diphones rather than phones, and falling back to halfphones where diphones are missing; avoiding signal processing; trying to select long chunks; and trying to limit audible breaks, notably with respect to F0.

A supporting factor was probably the fact that we used non-expressive speech material only. We have also built a test version of the system based on all speech material, but the quality seemed less reliable – when occasionally an expressive unit was chosen, it was negatively prominent compared to the rest of the sentence. Unsurprisingly, thus, it would seem that expressivity, if present, should be controlled rather than being present as an uncontrolled factor.

For Mandarin Chinese, we have managed to create a relatively intelligible voice that sounds reasonably natural in a language none of the authors speaks. This confirms that our new language support works in principle. It is possible that fine-tuning the cost weights could have improved the quality; however, this cannot be done without understanding the language.

As a matter of fact, we consider it an important downside of our approach that the cost weights need to be tuned manually at the moment. The tuning of weights constitutes an artistic or intuitive rather than scientifically well-founded element at a crucial point in the creation of a voice. Indeed, the quality changes drastically depending on the weights chosen. We currently set the weights for the target and join costs. by trial and error, using 10-20 example sentences (e.g., from recent news or from a novel). We adapt the relative importance of acoustic target costs (duration and F0), F0 join costs, and join vs. target costs, trying to maximise the resulting average length of consecutive chunks. At the same time we listen for clearly worse or better versions. Often however the differences are not fully clear, so that the ultimate decision is rather ad hoc. It would be preferable to have a scientifically well-founded approach to optimising the weights, which however would require a much more systematic investigation of the relation between target and join costs on the one hand and perceptual ratings on the other.

5. Conclusions

This paper has described the version of the MARY TTS system that was used in the Blizzard Challenge 2009. We reached good results in the English language section by using only neutral speech material for building the full voice.

We have outlined the new language support for MARY TTS, which will be made publicly available soon. Using this toolkit, we have been able to build a decent system supporting Mandarin Chinese, without actually speaking the language ourselves. However, as long as at least the adjustment of weights in the unit selection depends on manual tuning, it is questionable whether we will be able to substantially improve the quality of such a system. Given the copyright restrictions of the Blizzard data, we cannot make the Chinese voice publicly available.

A simple signal processing approach to formant enhancement for improving intelligibility in a telephone-based version of the challenge was not successful. Apparently the quality would have been better without the modifications.

The use of non-expressive speech material for building the full English voice, rather than predominantly expressive material in the previous year, has yielded good results. This indicates that for neutral target material, the use of expressive speech may not be appropriate. This finding seems intuitively plausible.

However, for expressive target styles, it is clear that suitably expressive material would be required. Such material being usually more variable with respect to, notably, pitch movements (except for low-arousal target styles such as depressed or bored speech), it would be more challenging to avoid discontinuities at unit join points.

6. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements n° 211486 (SEMAINE) and 231287 (SSPNet).

7. References

[1] M. Schröder, A. Hunecke, and S. Krstulović, "OpenMary ― Open Source Unit Selection as the Basis for Research on Expressive Synthesis," *Proc. Blizzard Challenge'06*, 2006.

[2] M. Schröder and A. Hunecke, "MARY TTS participation in the Blizzard Challenge 2007," *Proc. Blizzard Challenge 2007*, Bonn, Germany: 2007.

[3] M. Schröder et al., "The MARY TTS entry in the Blizzard Challenge 2008," *Proc. Blizzard Challenge*, Brisbane, Australia: 2008.
[4] M. Schröder and J. Trouvain, "The German text-to-speech synthesis system MARY: A tool for research, development and teaching," *International Journal of Speech Technology*, vol. 6, 2003, p. 365—377.
[5] A.W. Black and K. Lenzo, Festvox: Building synthetic voices, Edition 1.6, Carnegie Mellon University, PA, USA: 2002; http://www.festvox.org.
[6] L. M. Arslan, 1999, "Speaker Transformation Algorithm using Segmental Codebooks", Speech Communication, 28, pp. 211-226.
[7] J. R. Crosmer, 1985, Very Low Bit Rate Speech Coding Using the Line Spectrum Pair Transformation of the LPC Coefficients, Ph.D. Dissertation, Elec. Eng., Georgia Inst. Technology.
[8] "Mandarin Chinese Phonetics", *Third Edition*, http://www.zein.se/patrick/chinen8p.html

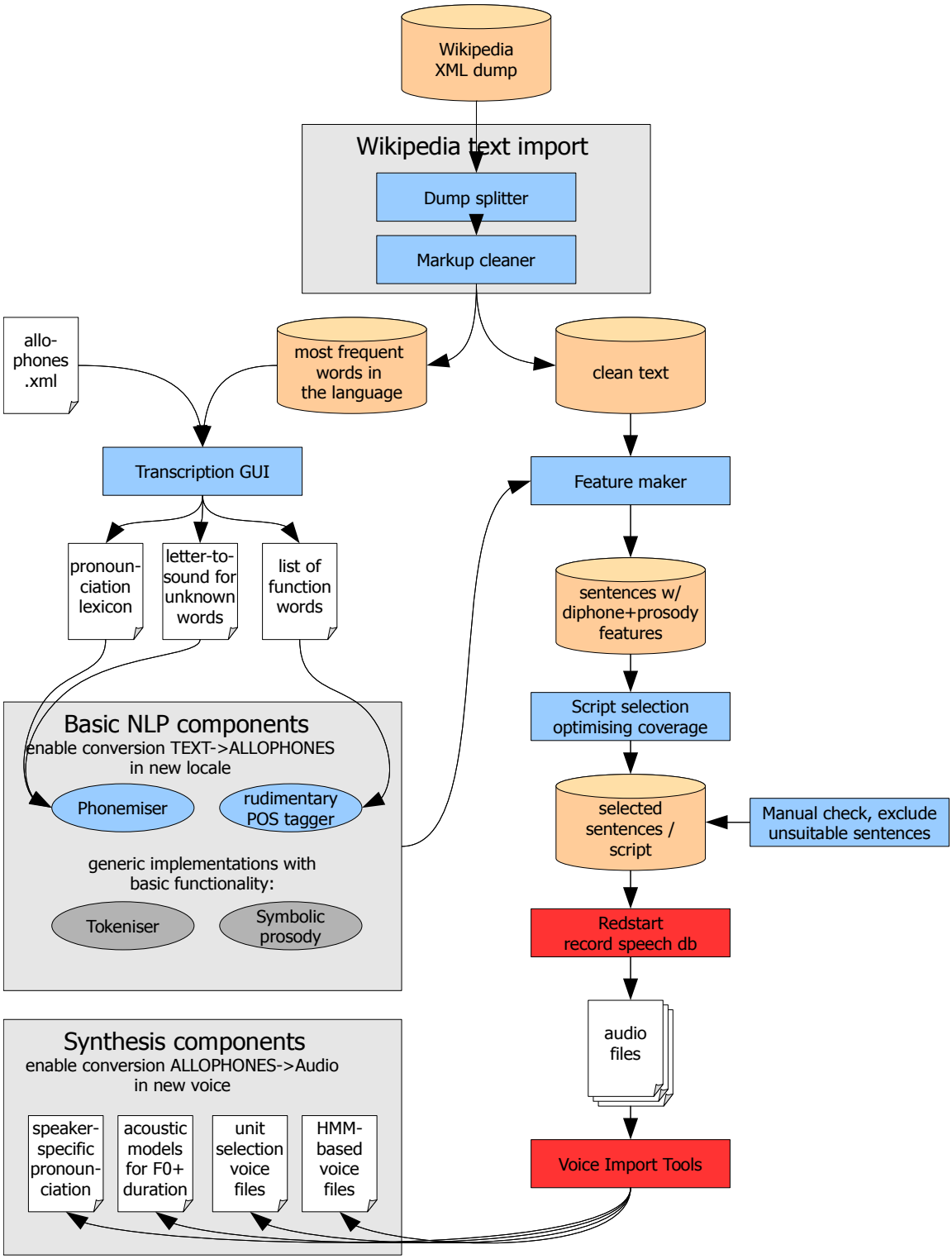

Figure 1: Workflow for supporting a new language in MARY TTS