The UPC TTS System Description for the 2008 Blizzard Challenge

Antonio Bonafonte[1], *Asunción Moreno*[1], *Jordi Adell*[1], *Pablo D. Agüero*[2], *Eleftherios Banos*[1], *Daniel Erro*[1], *Ignasi Esquerra*[1], *Javier Pérez*[1], *Tatyana Polyakova*[1]

[1]TALP Research Center, Universitat Politècnica de Catalunya, Spain
[2]Communications Lab, University of Mar del Plata, Argentina
pdaguero@fi.mdp.edu.ar

Abstract

This paper presents the UPC TTS system named Ogmios. It was used to generate the voices in UK English and Mandarin Chinese for Blizzard Challenge 2008. Ogmios is a system based on unit-selection using acoustic and phonetic features both in target and concatenation costs. Most of the modules of Ogmios rely on data driven techniques. This evaluation confirms that this framework allows fast development of synthetic voices in new languages that were not previously covered by the TTS: Mandarin Chinese.

The work in the TTS was devoted to prepare Ogmios to synthesize Mandarin Chinese, taking into account some particularities of the language regarding its writing and intonation. In the case of UK English we put the focus on rhythm, pauses and phonetic transcription.

The evaluators scored the UK English voices in an average of 3. High WER indicates intelligibility problems. Mandarin Chinese voice was scored with a lower score, near 2. Several measures indicate intelligibility and quality problems in this synthetic voice.

Index Terms: speech synthesis, synthesis systems, Blizzard evaluation.

1. Introduction

The objective of the evaluation of the 2008 Blizzard Challenge Initiative is to compare TTS systems at international level. The goal of Blizzard Challenge 2008 was to improve the quality and intelligibility of the synthesized speech. This year, the Centre for Speech Technology Research (CSTR) released a 15 hours UK English database and the National Laboratory of Pattern Recognition -Institute of Automation- of the Chinese Academy of Sciences, released a 6.5 hours Mandarin Chinese database of a female speaker (Beijing dialect).

The participants were asked to generate synthetic sentences using 3 voices:

- Voice A: from the full UK English database (about 15 hours)
- Voice B: from the ARCTIC subset of the UK English database (about 1 hour)
- Voice C: from the full Mandarin database (about 6.5 hours)

The UPC speech synthesis team participated in the 2008 Blizzard Challenge Initiative. This paper describes Ogmios, the UPC Text-to-Speech system used for the evaluation. The system was designed to cope with Spanish, Catalan and English languages [1] and, for the Blizzard Challenge 2008 its features were extended to cope with Mandarin Chinese. This paper is organised as follows: Section 2 describes the system, Section 3 shows some experiments with Mandarin Chinese intonation, Section 4 explains the process of building the voices, and finally, Section 5 presents and discusses the results of the evaluation.

2. System Description

2.1. Text and Phonetic Analysis

The first task of the system is to detect the structure of the document and to transform the input text into words. For this task we have used rules for tokenizing and classifying *non-standard words* in English, which are very similar to those used for Spanish and Catalan. The rules for expanding each token into *words* are language dependent, but are based in a few simple functions (spellings, natural numbers, dates, etc.) by means of regular expressions.

The second process is the POS tagger. Ogmios includes a statistical tagger based on FreeLing. The FreeLing package consists of a library providing language analysis services. Main services used of FreeLing library are PoS tagging and probabilistic prediction of unknown word categories. Freeling provides services for all currently supported languages: Spanish, Catalan, Galician, Italian, and English [2].

The input text for Mandarin Chinese was the Pinyin transcription of the utterance, initial/final data and POS tags. Therefore, it was not necessary any pre-processing for this language. The input text format provided by the organizers was converted to a markup format used at UPC, which is Ogmios compatible for enriched input.

2.1.1. Phonetic Transcription

The goal of the *phonetic* module is to provide the pronunciation of the words. This is used not only for producing the test sentences but also for transcribing the training database which is used for building the voices.

For UK English voices the pronunciation of each word is based on the Unisyn dictionary, provided by the University of Edinburgh [3]. It consists of 117K word entries. After listening to some samples, the accent chosen for this task is the RP accent. SAMPA was selected as the phoneset.

A finite state transducer (FST) was inferred to follow the Unisyn RP convention. The FST-based G2P [4] was trained using only the Unisyn dictionary. The performance of this method is around 74% correct for all words (94.15% phonemes are correct, with 4.43% of substitutions, 0.18% of insertions and 1.24% of deletions)

Some rules were hand-coded to model the pronunciation changes produced in continuous speech. For function words, a set of rules was produced based on factors like word's position in the sentence, part-of-speech and phrase accent. In continuous speech the function words usually lose their accented form and the full vowels are reduced to the shorter vowels or schwa. Furthermore, a set of phonotactic hand-crafted rules was applied. These rules cover different phenomena, from aspirated plosives to consonant assimilation and elision. In the training phase, the rules provided several pronunciation hypotheses which were considered by the segmentation process (see Section 4).

The input utterances for Mandarin Chinese TTS are provided using Pinyin, which is the most common standard for representing Standard Mandarin in the Latin alphabet. Since there is no official SAMPA symbol set defined for Mandarin Chinese at Sampa's website[1], a phoneme set called SAMPA-C [5] has been adopted. It is widely accepted as an accurate phoneme set for Mandarin, including some dialects. The total number of phonemes is 51, including 23 consonants, 17 vowels, 3 semi-vowels and 8 retroflexed finals.

According to the pronunciation of SAMPA-C symbol set, there is a mapping between Pinyin and SAMPA-C symbols. Therefore, a searching procedure over the Pinyin to SAMPA-C table is performed to obtain the phonetic transcriptions given the input utterance transcribed using Pinyin. The Pinyin transcription includes tone information, which is used as a syllable feature in the text-to-speech synthesis process.

In Chinese each syllable has a default tone when uttered in isolation. However, in continuous speech, the actually pronounced tone of a character may differ from the default one due to an effect called "tone sandhi". Tone sandhi denotes a set of rules about how to modify the tone depending on the syllable before or after. In our work we implemented these rules of the tone changes in the automatic training of prosodic modules, as shown in Sections 2.2.4 and 3.

2.2. Prosody

Prosody generation is done by a set of modules that sequentially perform all the tasks involved in prosody modelling: phrasing, duration, intensity and intonation. For the preparation of the Blizzard voices, a reduced database obtained after pruning the whole database was used (see section 4). For each of the three voices (A, B and C), we independently determined the maximum number of phoneme identification errors allowed per sentence. The files containing a larger number of errors were discarded. This threshold was automatically set based on the mean and standard deviation of the number of errors per file, so that approximately 85% of each data set was used during prosody estimation.

2.2.1. Phrasing

Phrasing is one of the key topics in the linguistic part of text-to-speech technologies and consists of breaking long sentences into smaller prosodic phrases. Boundaries are acoustically characterised by a pause, a tonal change, and/or a lengthening of the last syllable. Phrase breaks have strong influence on naturalness, intelligibility and even meaning of sentences.

In Ogmios phrasing is obtained using two algorithms. The first algorithm consists in a Finite State Transducer that translates the sequence of part-of-speech tags of the sentence into a sequence of tags with two possible values: break or non-break [6]. This uses the same tool which was used for the grapheme-to-phoneme task: x-grams [7]. The method uses very few features, but the results are comparable to CART using more explicit features.

The second algorithm predicts phrase break boundaries combining a language model of phrase breaks ($P(j_i|j_{i-k,i-1})$) [8] and probabilities of phrase breaks given contextual features ($P(j_i|C_i)$) [9]. Phrase break boundaries are found by maximizing the following equation:

$$J(C_{1,n}) = argmax_{j_{1,n}} \prod_{i=1}^{n} \frac{P(j_i|C_i)}{P(j_i)} P(j_i|j_{i-k,i-1}) \qquad (1)$$

The latest algorithm was chosen in this evaluation for UK English and Mandarin Chinese due to its better subjective performance in training data.

2.2.2. Duration

Phone duration strongly depends on the rhythmic structure of the language. For example, English is stressed-timed while Spanish is syllable-timed. Ogmios predicts phone duration with a two steps algorithm: prediction of the suprasegmental duration (syllable or stress unit), and then phone duration is predicted by factoring the suprasegmental duration.

The suprasegmental duration is predicted using CART. Features include the structure of the unit, represented by articulatory information of each phoneme contained in it (phone identity, voicing, point, manner, vowel or consonant), stress, its position in the sentence and inside the intonation phrase, etc.

Once the duration of the suprasegmental unit is calculated, the duration of each phoneme is obtained using a set of factors to distribute suprasegmental duration over its constituent phonemes. These factors are predicted using CART with a set of features extracted from the text, such as articulatory information of the phoneme itself and the preceding and succeeding ones, position in the unit, in the word and in the sentence, stress, and whether the unit is pre-pausal.

The duration model for UK English was stress-timed while Mandarin Chinese was syllable timed.

2.2.3. Intensity

The intensity of the phonemes is predicted by means of a CART. Features are again articulatory information of the actual, preceding and succeeding phone, stress, and the position in the sentence relative to punctuation and phrase breaks.

2.2.4. Intonation

Ogmios has two available intonation models: a superpositional polynomial model trained using JEMA (*Join feature Extraction and Modelling Approach* [10]), and a *f0 contour selection* model. In some cases, using the superpositional approach results in over-smoothed intonation contours with a loss of expressiveness.

Thus, in this evaluation we generate the f0 contour using the selection approach [11]. For each accent group we select a real contour from the database taking into account the *target cost* (position in the sentence, syllabic structure, etc.) and the *concatenation cost* (continuity). The selected contour is represented using a 4th order Bezier polynomial. The contour is generated using this polynomial, once the time scale is adapted to the required durations. The final result is a more expressive intonation contour than the JEMA model. However, in some

[1] http://www.phon.ucl.ac.uk/home/sampa/home.htm

cases, the contour is not adequate for the target sentence due to natural language understanding limitations of TTS systems.

2.3. Speech Synthesis

Our unit selection system runs a Viterbi algorithm in order to find the sequence of units $u_1 \ldots u_n$ from the inventory that minimises a cost function with respect to the target values $t_1 \ldots t_n$. The function is composed by a target and a concatenation cost: both of them are computed as a weighted sum of individual subcosts as shown below:

$$C(t_1 \ldots t_n, u_1 \ldots u_n) = w^t \sum_{i=1}^{n} \left(\sum_{m=1}^{M^t} w_m^t C_m^t(t_i, u_i) \right) + w^c \sum_{i=1}^{n-1} \left(\sum_{m=1}^{M^c} w_m^c C_m^c(u_i, u_{i+1}) \right)$$

where w^t and w^c are the weights of the global target and concatenation costs ($w^t + w^c = 1$); M^t is the number of the target sub-costs and M^c the number of concatenation sub-costs; $C_m^t(.)$ is the m th target sub-cost which is weighted by parameter w_m^t; and $C_m^c(.)$ is the m th concatenation sub-cost weighted by w_m^c.

Tables 1 and 2 show the features used for defining the subcost functions. There are two types of sub-costs functions. Binary, which can only have 0 or 1 values, and continuous. For continuous sub-costs functions, a distance function is defined and a sigmoid function is applied in order to restrict their range to $[0 - 1]$.

To adjust the target weights, we applied a similar approach to the one proposed in [12]. For each pair of units, we compute their distance using feature vector (MFCC, f0, energy) taken every 5 msec. Let $\overline{d}$ be the vector of all distances for each pair of units, C a matrix where $C(i, j)$ is sub-cost j for unit pair i and $\overline{w}$ the vector of all weights to be computed. If we assume $C\overline{w} = \overline{d}$ then it is possible to compute $\overline{w}$ as a linear regression. In other words, the target function cost becomes a linear estimation of the acoustic distance. The weights of the concatenation subcosts functions were adjusted manually.

phonetic accent	B
duration difference	C
energy difference	C
pitch difference	C
pitch diff. at sentence end	C
pitch derivative difference	C
pitch deviate sign is different	B
accent group position	B
triphone	B
word	B

Table 1: Target costs: B stands for binary cost and C for continuous cost.

Concerning the waveform generation process, in our experience, listeners assign higher quality scores to the synthetic utterances where the prosodic modifications are minimal. Thus, most of the units selected for generating synthetic speech are simply concatenated using glottal closure instant information, without any prosodic manipulation. Therefore, the use of the information provided by the prosody generation block is restricted to the unit selection process.

energy	C
pitch	C
pitch at sentence end	C
spectral distance at boundary	C
voice-unvoiced concatenation	B

Table 2: Concatenation costs: B stands for binary cost and C for continuous cost.

3. Intonation experiments with Mandarin Chinese

Our work with Mandarin Chinese for Blizzard Challenge 2008 included the study of JEMA [10] and its use to generate the fundamental frequency contour. The experiments were performed using the Mandarin Chinese database provided by the The National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, It consists in 6.5 hours of utterances in Mandarin Chinese (Beijing dialect) of a female speaker. The utterances are transcribed using Chinese Traditional Characters and Pinyin. Part-of-Speech tags are provided as well as complementary information to help in the task of prosody modelling in Blizzard 2008 [13].

The intonation model was trained using a set of features extracted from the transcription of the utterances. Due to the superpositional approach, two prosodic units are used: syllable and phrase. The syllable intonation contours are predicted based on: position of the syllable relative to the word and phrase, syllable tone, preceding and succeeding tones (in order to deduce rules such as tone Sandhi), prepausal and the Pinyin transcription of the syllable. The phrase component is predicted using the following features, number of words and syllables in the phrase, punctuation, POS preceding and succeeding the break and POS sequence in the phrase.

The experiments are conducted to study the naturalness and quality of the intonation compared with natural speech. The training data consists of 70% of the database and the other 30% is test data. Two objective measures are used to study the difference between real and predicted contours: root mean squared error (RMSE) and correlation coefficient.

The RMSE for training and test data was $37.4Hz$ and $37.5Hz$, and the correlation was 0.824 and 0.833, respectively. Such high values for correlation show that the intonation model achieves a high resemblance in trajectory with a natural pitch contour. An intonation model trained using SEMA (Separate Extraction and Modelling Approach: parameter extraction and model training in separate steps) was also included in the experiments. The resulting RMSE and correlation were $38.5Hz$ and 0.82.

The difference between natural (reference) and predicted pitch contours shown by RMSE are due to two facts. On the one hand, speakers with higher pitch range will show higher RMSE in the linear scale of frequency, as shown in many papers in the literature. On the other hand, although the high correlation shows an appropriate trajectory in pitch contour, the predicted contours may not have the full pitch range of the original speaker due to the inherent smoothing of this kind of clustering approaches. The representative pitch contour of the cluster may not have enough excursion to match the natural contour, as shown in Figure 1.

The improvement obtained by the JEMA approach compared with SEMA is small due to the high influence of syllable tone in the pitch contour, which partially masks the melioration

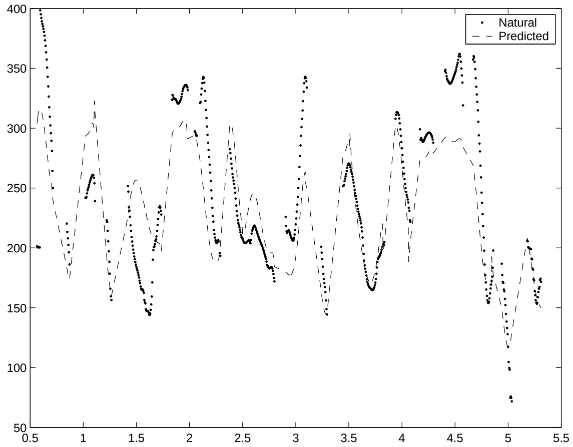

Figure 1: Difference in excursion between natural and predicted pitch contours.

of the superpositional approach. This statement is supported by the study of the relevance of the features. It reveals that tone features are the most relevant for the pitch contour of the syllable: syllable tone, and preceding and succeeding tones. The relevant features for phrase component are the number of syllables and words in the phrase.

A subjective experiment was also performed to study the quality and naturalness from the point of view of perception. The listening test was done by 88 native mandarin Chinese speakers. They were asked to score natural and synthetic utterances in a five-point scale. The synthetic audios were obtained using resynthesis with Praat [14].

The natural utterances were scored as 4.85 and 4.72 for quality and naturalness, respectively. Meanwhile,. predicted contours obtained for quality 4.33 and for naturalness 4.06. The high quality shows the little distortion introduced by PSOLA manipulation in resynthesis. The score in naturalness for predicted contours is explained by the limitations in excursion previously explained.

Despite the good experimental results obtained with JEMA, we decided to use the f0 contour selection approach (Section 2.2.4). JEMA has a smoothing effect on the intonation curve that may lower the expressiveness of synthetic speech.

4. Building the Blizzard Voices

Once the normalization and phonetic transcription rules are ready (section 2.1), our system is able to build a new voice automatically from the audio files and their corresponding prompts. This automatic procedure consists of four main steps: automatic segmentation of the database, training of the prosodic models, selection weights adjust plus database indexing. The prosody training and the selection weights adjust procedures have been described in previous sections. Therefore, in the present section, we will describe the segmentation process and the database indexing.

Once the database was supplied we built the unit inventory. In our system, the units are context dependent demiphones. However, the selection algorithm forces the use of diphones imposing a high cost in phone transitions. The database is automatically segmented into phones by means of the HMM-based aligner named Ramses [15]. We used the front-end described in section 2.1 to automatically transcribe the whole database into phones.

Afterwards, we trained a different set of context-dependent demiphone HMM models from each data set, corresponding to each of the three voices. The phone boundaries are determined using a forced alignment between the speech signal and the models defined by the phonetic transcription. A silence model, trained at punctuation marks, was optionally inserted at each word boundary during the alignment. In addition, the detected silences are also used for the pause prediction model (see Section 2.2).

Previous experiments have shown that when a correct phonetic transcription is given, HMM models can achieve similar speech synthesis quality than manual segmentation [16, 17]. Therefore, additional effort was devoted to phonetic transcription and database pruning to obtain correctly segmented voices, as show in the following paragraphs.

Automatic phonetic transcription of a speech synthesis database has to cope with pronunciation variants, pronunciation errors and recording noise. In order to overcome the former problem, the alignment took into account all possible transcriptions of a single word. At this point, the alignment may have errors either because there is a mismatch between front-end and speaker production or because there is an alignment error.

We assume that wrong units will never represent a big portion of the database and that it is affordable to reject such part of it. Therefore we tried to detect undesired units in order to remove them from the inventory by means of a pruning procedure. After computing the alignment likelihood for every unit, 10% of them, those with worst scores, were removed. Previous experiments have shown that it is possible to remove 90% of wrong units by means of this pruning procedure [18].

Once the speech signals were segmented and the list of sentences are ready, we can start building the voices for our TTS system. The process consists of three main steps: feature extraction, unit indexing and voice generation. The first step extracts F0, duration, energy and MFCC for each speech unit. The index file contains the relevant information needed for computing the target and concatenation costs. In the last step, the parameters of the prosody models and the weights of the unit selection algorithm are computed.

5. Results

The organizers of Blizzard asked to synthesize 970 utterances for UK English voices: full Roger and ARCTIC. The test sentences came from various sources: Blizzard 2007 sentences, newspaper sentences, conversational sentences [19][20], semantically unpredictable sentences and emphasis sentences.

Participants with synthesizers in Mandarin Chinese were asked to generate one voice, and they synthesized 967 utterances from two main sources: news sentences and semantically unpredictable sentences.

5.1. Evaluation framework

The subjective evaluation was conducted online. For English, on registration, listeners were assigned to hear voices built with one of the datasets. For Mandarin, there was just one dataset.

The subjective evaluation was divided into five sections, as explained in the README file of the evaluation results:

- **Section 1**: The listeners chose a response that represented how similar the synthetic voice sounded to the voice in the 4 reference samples of the original speaker on a scale from 1 (sounds like a totally different person) to 5 (sounds like exactly the same person).

- **Section 2**: The evaluators chose whether the two sentences of two participating systems were similar or different in terms of their overall naturalness.
- **Section 3**: Mean Opinion Score (MOS) in the news domain.
- **Section 4**: MOS in novel domain for English and in the news domain for Mandarin Chinese. In each part of section 3 and 4 listeners listened to one sample and chose a score which represented how natural or unnatural the sentence sounded on a scale of 1 (completely unnatural) to 5 (completely natural).
- **Section 5**: Semantically Unpredictable Sentences (SUS) designed to test the intelligibility of the synthetic speech. Listeners heard one utterance in each part and typed in what they heard. The results are expressed as WER (Word Error Rate).

The procedure for calculation of error rates in Mandarin Chinese was:

1. Convert any traditional Chinese characters to simplified Chinese characters.
2. Calculate Character Error Rate (CER) using a similar procedure to WER, treating each character as a word. No spelling correction was used.
3. Convert each character to Pinyin+Tone (a one-to-many mapping); the result is a lattice of possible Pinyin+Tone sequences.
4. Calculate Pinyin+Tone Error Rate (PTER), choosing the Pinyin+Tone path through the lattice that gives the lowest PTER.
5. Strip the tones leaving only Pinyin, and calculate Pinyin Error Rate (PER), choosing the Pinyin path through the lattice that gives the lowest PER.

5.2. Analysis of the results

Table 3 shows the results for similarity, MOS and WER for voice A, individualized by section and evaluator class: EE (paid UK students), EI (paid Indian students), ER (volunteers) and ES (speech experts). Table 4 shows the results for voice B, with a degradation with respect to voice A due to the smaller data set.

Sections	**EE**	**EI**	**ER**	**EE**
Section (Sim) 1	3.38	2.76	2.96	3.01
Section (MOS) 3	2.57	3.21	2.81	3
Section (MOS) 4	2.76	3.14	2.49	3.01
Section (WER) 5	0.35	0.54	0.69	0.53

Table 3: Similarity, MOS and WER calculated by evaluator and section for voice A.

The similarity of the synthesized speech with the original speaker is medium, and the MOS is also in half the scale. The main reason for that is originated in articulation problems in the output speech, as shown by WER.

One of the main problems found during the process of voice generation was the low probability found in some words and the difficulty to generate appropriate phonetic transcriptions for voice segmentation and speech synthesis. It is necessary to obtain correct phonetic transcriptions given text prompts in both cases, because dialect is an important part of the identity. As a consequence, it is not possible to use multiple word transcriptions for a given word to improve segmentation without considering that information during the grapheme to phoneme conversion. Such consistency will be considered in future works.

The analysis of the global results (merged for all speakers) shows a small difference in the MOS scale between voices A (MOS=2.9) and B (MOS=2.7). As a consequence, Ogmios does not achieve a high gain due to the higher amount of data. The same happens with similarity (3.05 for voice A and 2.8 for voice B) and WER (0.418 for voice A and 0.445 for voice B).

During the analysis of the results for UK voices appeared differences between evaluators in WER (voice A: 0.26 for EE and 0.49 for ES;voice B: 0.26 for EE and 0.5 for ES). They remain unexplainable for us.

Sections	**EE**	**EI**	**ER**	**EE**
Section 1 (Sim)	2.89	3	2.69	3.25
Section 3 (MOS)	2.52	2.60	2.43	2.72
Section 4 (MOS)	2.63	2.97	2.58	3.02
Section 5 (WER)	0.40	0.65	0.73	0.68

Table 4: Similarity, MOS and WER calculated by evaluator and section for voice B.

Table 5 shows the subjective results for Mandarin Chinese. All scores are very low, showing strong problems to generate a synthetic voice for Ogmios in this language.

The development of the system in Chinese was very difficult due to the missing experience in oriental languages and the small feedback provided by a non-expert Chinese student, which could not offer enough information to correct general problems.

Machine learning techniques showed that the performance with objective measures, such as RMSE and correlation measures, were in the range of value of other languages, except in the case of pauses.

Our work with word segmentation, Pinyin transcription of OOV words and POS annotation did not achieve good results at the beginning of voice development. However, this problem was solved by the organizers providing Pinyin transcriptions and POS tags as part of the available information for training and test.

Although the overall performance of the synthesis system is not yet close to state-of-the-art systems, we have obtained very good results with our experiments in Mandarin Chinese intonation modelling [21], as shown in Section 3.

Sections	**MC**	**ME**	**MR**	**MS**
Section 1 (Sim)	2.27	1.95	1.91	2.08
Section 3 (MOS)	2.2	1.68	1.45	1.70
Section 4 (MOS)	2.13	1.48	1.34	1.47
Section 5 (CER)	0.57	0.46	0.63	0.80
Section 5 (PTER)	0.43	0.20	0.43	0.72
Section 5 (PER)	0.51	0.36	0.56	0.78

Table 5: Similarity, MOS, CER, PTER and PER calculated by evaluator and section for voice C.

6. Conclusions

This paper describes Ogmios, the text-to-speech system developed at UPC. Ogmios has been designed to be multilingual, but till now, most of our efforts addressed the Spanish and Catalan languages. The Blizzard Challenge experience has shown us that we are able to build a new voice, in a new language, with a limited amount of work.

However, the results in English and Mandarin Chinese are significantly worse that the results obtained in Spanish: MOS close to 4 [1]. We believe that the reason for this gap is not related with technological limitations of our system working in English or Mandarin Chinese, but with the difficulties for tuning the system by non-native speakers.

One important aspect that will be addressed in future evaluations is the search for an automatic procedure to tune the grapheme-to-phoneme conversion to match speaker dialect and style. Low probability scores of HMM models for some automatic phonetic transcription (after the use of English weak forms and phonotactic rules, or Pinyin to SAMPA-C) show problems that need to be addressed in the future.

In this we also prove that although most of the components in the text-to-speech systems achieved good performances (intonation, duration, intensity, phonetic transcription in English) the overall results are medium for UK English and low for Mandarin Chinese due to the strong interaction between components.

We encourage the organisers to continue with this challenge and we support their idea of including other languages in the evaluation and we offer our Catalan resources for next evaluation rounds.

7. Acknowledgements

This work has been funded by the Spanish Government (project AVIVAVOZ, TEC2006-13694-C03).

The authors thank to The Blizzard 2008 Team for organising, and the Centre for Speech Technology Research (CSTR) and the National Laboratory of Pattern Recognition -Institute of Automation- of the Chinese Academy of Sciences, for providing the speech data.

8. References

[1] Bonafonte, A., Agüero, P. D., Adell, J., Perez, J., and Moreno, A., “Ogmios: The UPC text-to-speech synthesis system for spoken translation”, Proceedings of TC-STAR Workshop, Barcelona, Spain, June, 2006.

[2] Atserias, J., Casas, B., Comelles, E., Gonzalez, M., Padro, L., and Padro, M., “FreeLing 1.3: Syntactic and semantic services in an open-source NLP library”, Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006), ELRA. Genoa, Italy. May, 2006.

[3] Fitt, S., “Documentation and User Guide to Unisyn Lexicon and Post-Lexical Rules”, Centre for Speech Technology Research, University of Edinburgh, 2000.

[4] Galescu, L., and Allen, J., “Bi-directional conversion between graphemes and phonemes using a joint n-gram model”, Proceedings of the 4th ISCA Speech Synthesis Workshop, Perthshire, Scotland, September, 2001.

[5] Xiaoxia, C., Aijun, L., Guohua, S., Wu, H., and Zhigang, Y., “An Application of SAMPA-C for Standard Chinese”, Proceedings of the 6th International Conference on Spoken Language Processing, Beijing, China, October, 2000.

[6] Bonafonte, A., and Agüero, P. D., “Phrase break prediction using a finite state transducer”, Proceedings of the 11th International Workshop on Advances in Speech Technology, Maribor, Slovenia, July, 2004.

[7] Bonafonte, A., “Language modeling using x-grams”, Proceedings of International Conference on Spoken Language Processing, 1996.

[8] Black, A., and Taylor, P., “Assigning Phrase Breaks from Part-of-Speech Sequences”, Proceedings of Eurospeech, 1997.

[9] Agüero, P. D., and Bonafonte, A., “ Phrase break prediction: a comparative study”, XIX Congreso de la Sociedad Española para el procesamiento del Lenguaje Natural. Alcala de Henares, Spain, September, 2003.

[10] Agüero, P. D. and Bonafonte, A., “Intonation Modeling for TTS Using a Joint Extraction and Prediction Approach,”, Proceedings of the International Workshop on Speech Synthesis, Pittsburgh, USA, 67-72, 2004.

[11] Malfrère, F., Dutoit, T., and and Mertens, P., “Automatic prosody generation using suprasegmental unit selection”, Proceeding of the 3rd ISCA Speech Synthesis Workshop, Jenolan Caves, Australia, December, 1998.

[12] Hunt, A., and Black, A., “Unit selection in a concatenative speech synthesis system using a large speech database”, Proceedings of ICASSP, Atlanta, Georgia, 1996.

[13] Black, A. W., King, S., and Tokuda, K., “The Blizzard Challenge 2008 - Evaluating Corpus based Speech Synthesis on Common Databases”, http://www.synsig.org/index.php/Blizzard Challenge 2008.

[14] Lemon, O., “Praat: Doing Phonetics by Computer”, http://www.fon.hum.uva.nl/praat/.

[15] Bonafonte, A., Mariño, J. B., Nogeuiras, A., and Rodriguez Fonollosa, J. A., “RAMSES: el sistema de reconocimiento del habla continua y gran vocabulario desarrollado por la UPC”, VIII Jornadas de Telecom I+D (TELECOM I+D ’98), Madrid, Spain, October, 1998.

[16] Makashar, M. J., Wightman, C. W., Syrdal, A. K., and Conkie, A., “Preceptual evaluation of automatic segmentation in Text-to-Speech synthesis”, Proceedings of ICSLP, Beijin, China, October, 2000.

[17] Adell, J., Bonafonte, A., Gómez, J. A., and Castro, M. J., “Comparative study of automatic phone segmentation methods for TTS”, Proceedings of ICASSP, Philadelphia, PA, USA, March, 2005.

[18] Adell, J., Agüero, P. D., and Bonafonte, A., “Database pruning for unsupervised building of text-to-speech voices”, Proceedings of ICASSP, vol. 1, Toulouse, France, May, 2006.

[19] Lemon, O., Georgila, K., and Henderson, J., “Evaluating Effectiveness and Portability of Reinforcement Learned Dialogue Strategies with real users: the TALK TownInfo Evaluation”, IEEE/ACL Spoken Language Technology, 2006.

[20] Boersma, P. and Weenink, D., “The TownInfo Spoken Language Understanding Corpus”, Edinburgh University, School of Informatics, 2008.

[21] Agüero, P. D., Bonafonte, A., Lu Yu, and Tulli, J.C., “ Intonation Modeling of Mandarin Chinese Using a Superpositional Approach,”, To appear in Proceedings of Interspeech, Brisbane, Australia, 2008.