The UPC TTS System Description for the 2007 Blizzard Challenge

Antonio Bonafonte, Jordi Adell, Pablo D. Agüero, Daniel Erro,
Ignasi Esquerra, Asunción Moreno, Javier Pérez, Tatyana Polyakova

TALP Research Center
Department of Signal Theory and Communications
Universitat Politècnica de Catalunya, Barcelona, Spain
antonio.bonafonte@upc.edu

Abstract

This paper presents the evaluation of *Ogmios*, the UPC TTS system carried out within the Blizzard Challenge Initiative, 2007. *Ogmios* is a unit-selection based system. Prosodic models are used to select the units using acoustic measures in the target cost but the selected units are not modified.

Most of the modules of *Ogmios* rely on data driven techniques. This evaluation confirms that this framework allow fast development of synthetic voices in new languages.

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

1. Introduction

Recently the UPC 's speech synthesis team has participated in the 2007 Blizzard Challenge Initiative, an evaluation campaign whose objective was to compare TTS systems at an international level. The global goal set in the framework of Blizzard challenge was to improve the quality and intelligibility of the synthesized speech. This year, ATR-SLC release an eight-hour American English speech database. The participants were asked to generate synthetic sentences using 3 voices: the *A* voice derived from the full data, the *B* voice, derived from the ARCTIC subset, and the *C* voice, derived by a subset of data defined by each participant.

This paper describes *Ogmios*, the UPC Text-to-Speech system used for the evaluation. The system was designed to cope with Spanish and Catalan languages [5] and, for the Blizzard Challenge its features were extended to cope with US English. The paper is organised as follows: In Section 2 we describe the system. Section 3 describes the process of building the voices. Specifically, the definition of the *C* voice and the segmentation of the speech database into phones. Finally, section 4 present and discuss 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 extended our system to English. The rules for tokenizing and classifying *non-standard words* 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.).

The second process is the POS tagger. *Ogmios* includes a basic statistical tagger. The n-gram statistics were estimated using 1 million of tokens from the WSJ Corpus using the Penn Tree bank POS system.

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.

The pronunciation of each word is based on the Unisyn dictionary, provided by the University of Edinburgh [7]. It consists of 110K word entries. After listening to some samples, the accent chosen for this task is the rhotic version of the NYC accent[1]. SAMPA was selected as the phoneset. Since we believed that it was better to have a previously validated transcription than let this problem be solved by the grapheme-to-phoneme (G2P) converter, the LC-STAR US-English dictionary [1] was *merged* with the system dictionary. This dictionary includes 50K common words and 50K North American proper names. Each proper name is marked with a label whether it is a geographic, person's or company name. In this context, *merging* means that only the words that were not found in the Unisyn dictionary were added from the new dictionary. Some small changes where done to adapt the phoneset of both SAMPA dictionaries. After the evaluation deadline, a deeper study of dictionary compatibility was performed. We discovered that these dictionaries are not compatible in their original format. The analysis of words which appear in both dictionaries show that only 30% of the entries provide the same phonetic transcription. After the evaluation a finite state transducer (FST) was inferred to transform the LC-STAR dictionary to follow the Unisyn NYC convention [12]. Basically, the phonetic transcription of the words that appears in both dictionaries were aligned and the mapping was estimated using the same FST-based tecnique than the one applied for G2P. The use of this technique raised the entries with the same transcription to 83%. If the word wasn't found in the *merged* dictionary then the grapheme-to-phoneme (G2P) conversion was necessary. The FST-based G2P [8] was trained using only the Unisyn dictionary. The performance of this method is around 70% correct for common words and 53% for proper names.

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

[1] One reviewer comments seems to indicate that the General American (GAM) accent should have been a better option for this voice

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 3.1).

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 3.1). For each of the three data sets (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 on 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 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 is the same tool which was used for the grapheme-to-phoneme task. The method uses very few features, but the results are comparable to CART using more explicit features [6].

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 syllable duration and prediction of phone duration.

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

Once the duration of the syllable is calculated, the duration of each phoneme is obtained using a set of factors to distribute syllable duration along its 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 syllable, in the word and in the sentence, stress, and whether the syllable is pre-pausal.

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* [4]), 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 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 3rd 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 cases, the contour is not natural for the target sentence.

2.3. Speech Synthesis

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

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

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.

Table 1 show the features used for defining the sub-cost 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 [9]. 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 dinstance. distance. The weights of the concatenation sub-costs functions were adjusted manually.

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

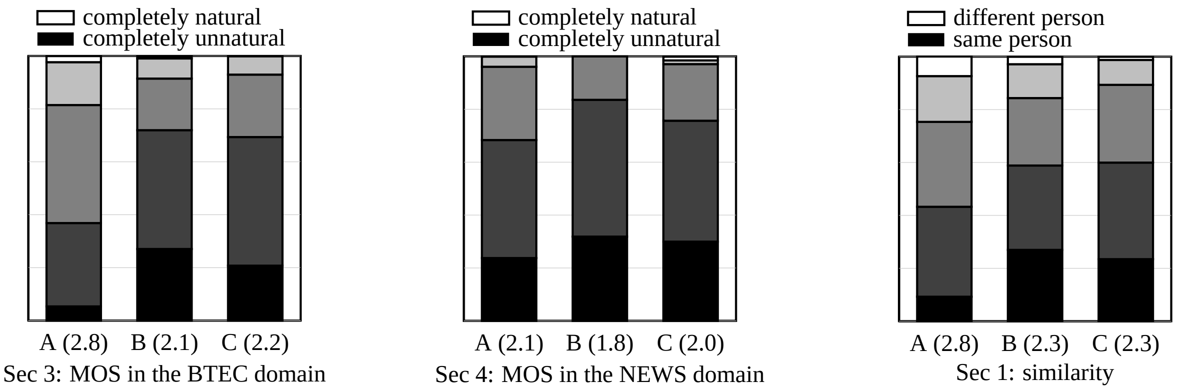

Figure 1: *Results for tests 3, 4 and 1 for the three UPC voices*

Target costs	
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

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

Table 1: Sub-costs summary plus their corresponding weights. B stands for binary cost and C for continuous cost.

3. 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 on 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, the sentence selection process needed for Voice C and the database indexing.

3.1. Segmentation of the Speech Database

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 a HMM-based aligner. 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 aligned 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 [10, 3]. Therefore, additional effort was phonetic transcription and database pruning.

Automatic phonetic transcription of a speech synthesis database has to cope with pronunciation variants and pronunciation errors or 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 not be 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. The alignment likelihood of each unit is computed and the 10% with worst values was removed. Previous experiments have shown that it is possible to remove 90% of wrong units by means of this pruning procedure [2].

3.2. The *C* voice

As stated in the Blizzard Challenge 2007 rules, *Voice C* has to be build from a subset of the full corpus, selected only from the text data without using any information from the speech signal. Phonetic or prosodic labelling, as well as any other process needed to train TTS modules, have to stand only in those sentences text. In addition, its total duration must be no more than 2914 seconds (i.e. ARCTIC subset duration).

The starting corpus for the selection was the totality of the available sentences, namely ARCTIC, BTEC and NEWS subsets. The first step was to reject sentences with many out-of-dictionary words, typically foreign words. This led to a 5223-sentence corpus to start with. These sentences were passed through the text analysis and phonetic transcription modules to represent the sentences by phonetic *units* including some prosodic features. Four different kind of units were considered: phones, context-phones, diphones and context-diphones. Context was defined as syllable accent (true or false) and word position in the intonation group. The sentences were selected with a greedy algorithm using *CorpusCrt*, our publicly available tool [13]. After some experiments, context-phones were chosen because they give more information than phones and there are enough repetitions of each to fulfill the criteria.

Sentences were chosen to maximize the number of repetitions per phonetic unit in the final corpus. Most of the context dependent phones have 10 or more repetitions in the corpus. Sentences between 10 to 150 units long were selected.

The resulting corpus has 727 sentences, and its composition is very close to the full Blizzard corpus. In particular, it is made of 17% from Arctic, 62% from Btec and 21% from News sets.

3.3. The Voice Building

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.

4. Results

For each of the 3 voices, the 400 test sentences were send to *Ogmios* and were evaluated by more than 100 human judges. The voices were rated in terms of mean opinion score (MOS). Figure 1 shows the results for three parts of the test. The left graph shows the MOS in the conversational domain, one column for each voice. For each column, we indicate the percentage of judges that rate the voice in each category, from completely unnatural (1, black) to completely natural (5, white). In brackets we present the mean value. As can be seen, the performance of voice B and voice C is very similar and significantly worse than voice A.

The center graph shows the results for the News domain. The results for this domain are significantly worse than for the conversational domain. This seems to be a general conclusion for all the systems and it seems to be related with lenght of the sentences which are significantly longer in the News domain. We also can observe how in this test, the performance of the voice A (full data) is quite similar to the performance of voice C (reduced data).

Finally, the right graph shows the results of the similarity test. We were expecting that our contatenative system, for all the three voices were very close to 1. However, the results are quite high, in particular for the best voice (A). We think that the judges are considering naturalness or other issues and not just a *segmental* interpretation of identity.

5. 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 are significantly worse that the results obtained in Spanish, where we obtained MOS close to 4 [5]. We believe that the reason for this gap is not related with technological limitations of our system working in English, but with the difficulties for tuning the system by non-native speakers.

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.

6. Acknowledgements

This work has been funded by the European Union (project TC-STAR, IST-2002-FP6-506738) and by the Spanish Government (project AVIVAVOZ, TEC2006-13694-C03).

The authors thank to The Blizzard 2007 Team for organising and this evaluation, and AT&R for providing the speech data.

7. References

[1] “Lexica and corpora for speech-to-speech translation components,” 2005. [Online]. Available: http://www.lc-star.com

[2] J. Adell, P. D. Agüero, and A. Bonafonte, “Database pruning for unsupervised building of text-to-speech voices,” in *Proc. of ICASSP*, vol. 1, Toulouse, France, May 2006, pp. 889–892.

[3] J. Adell, A. Bonafonte, J. A. Gómez, and M. J. Castro, “Comparative study of automatic phone segmentation methods for TTS,” in *Proc. of ICASSP*, Philadelphia, PA, USA, Mar. 2005.

[4] P. D. Agüero and A. Bonafonte, “Intonation modeling for TTS using a joint extraction and prediction approach,” in *Proc. of the 5th ISCA Speech Synthesis Workshop*, Pittsburgh, USA, June 2004.

[5] A. Bonafonte, P. D. Agüero, J. Adell, J. Pérez, and A. Moreno, “Ogmios: The UPC text-to-speech synthesis system for spoken translation,” in *Proc. of TC-Star Workshop*, Barcelona, Spain, June 2006.

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

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

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

[9] A. Hunt and A. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” in *Proc. of ICASSP*, vol. 1, 1996, pp. 373–376, Atlanta, Georgia.

[10] M. J. Makashay, C. W. Wightman, A. K. Syrdal, and A. Conkie, “Preceptual evaluation of automatic segmentation in Text-to-Speech synthesis,” in *Proc. of ICSLP*, Beijin, China, Oct. 2000.

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

[12] T. Polyakova and A. Bonafonte, “Fusion of dictionaries in voice creation and speech synthesis task,” in *Proc. of SPECOM*, Moscow, Russia, Oct. 2007.

[13] A. Sesma and A. Moreno, “Corpuscrt 1.0: Diseño de corpus orales equilibrados,” UPC, Tech. Rep., Dec. 2000.