Ogmios: the UPC entry for the Albayzin 2010 TTS Evaluation

Antonio Bonafonte[1], *Ignasi Esquerra*[1], *Asunción Moreno*[1], *Pablo Daniel Agüero*[2]

TALP Research Center, Universitat Politècnica de Catalunya, Spain [1]
Communications Lab, University of Mar del Plata, Argentina [2]

Abstract

This paper describes Ogmios, the UPC TTS system that was used in the 2010 Albayzin Evaluation. Ogmios is a concatenation system that builds the synthetic sentence from demiphones selected from the training database. In this evaluation round, the database was provided by the organization and it has been phonetically transcribed and segmented automatically using using the development tools included in Ogmios. Based on the segmentation the synthetic voice is build. Each voice includes the segmental inventory (waveforms of acoustic units), prosodic models (breaks, segmental duration and f0 prediction models) and the weights of the selection cost function. In order to have a better prosody models, in this evaluation external data has been used to ensure better prosodic coverage.

1. Introduction

This paper describes Ogmios, the UPC Text-to-Speech system used for the evaluation. The system was originally designed for Spanish and Catalan but has been extended to English and Mandarin [1, 2].
During the last 2 years, Ogmios has extended their functionalities: it is able to produce conversational speech to some degree [3] and to cope with multilingual text [4]. It also includes HTS as a possible back-end for speech synthesis. However, the system used in Albayzin'2010 is basically the same that was used in Albayzin'2008. We include here the description from [5] for completness. The main difference is that we have used prosody models estimated from other male speaker with a larger (10h) database. Our prosody algorithms, in particular, the phrassing component, did not perform well with the Albayzin'2010 database.
This paper is organised as follows: Section 2 describes the system and Section 3 explains the process of building the voices. Finally, section 4 describes 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 Spanish. 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 [6].

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 Spanish the pronunciation of each word is based on a set of rules that take into account the transcription rules of Spanish and phonotactics.
Some particular words are transcribed using a lexicon, specially foreign words, abbreviations and signs.

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.

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 [7]. This uses the same tool

which was used for the grapheme-to-phoneme task: x-grams [8]. 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 [9] and probabilities of phrase breaks given contextual features [10]. Phrase break boundaries are found by maximizing the following equation:

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

The latest algorithm was chosen in this evaluation for Spanish 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.

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* [11]), 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 [12]. 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 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 $u_1 \dots u_n$ 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.
Tables 1 and 2 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 [13]. 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 sub-costs 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.

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.

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.

3. Building the Albayzin voice

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 [14]. 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 [15, 16]. 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 [17].

In this evaluation we do not include any pruning due to the small amount of data provided to generate the synthetic voice. Therefore, we rely on spectral measures at unit selection to avoid problematic units.

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. Experimental Results and Discussion

All the voice models were estimated automatically from the voice using Ogmios training modules.

The similarity of the synthetic voice with the original one is relatively high, with mean value > 3. Even using prosodic models derived from other speaker, the listeners can perceive that the synthetic speaker is the same than the original voice. This should be a common feature to

any concatenative system. This question is more relevant for statistical or parametric methods.
With respect to naturalness, the MOS is relatively low, 2.60. We think that the reason is that even with models trained from data, the algorithms (segmentation, unit selection weights, signal processing) are tuned to larger databases. For smaller databases different tuning, or even different synthesis methods should be applied.

5. Acknowledgements

This work has been funded by the Spanish Government (project BUCEADOR: Busqueda de Informacion en Contenidos Audiovisuales Plurilingues TEC2009-14094-C04-01)
The authors thank to the Albayzin 2010 Team for organising the evaluation and providing the speech data.

6. 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] Bonafonte, A., Moreno, A., Adell, J., Agüero, P.D., Banos, E., Erro, D., Esquerra, I., Perez, J., and Polyakova, T., "The UPC TTS System Description for the 2008 Blizzard Challenge", Blizzard 2008, Brisbane, Australia, September, 2008.

[3] Adell, Jordi and Bonafonte, Antonio and Escudero, David, "Synthesis of Filled Pauses Based on a Disfluent Speech Model", Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),Dallas, USA, March, 2010

[4] Polyakova, Tatyana and Bonafonte, Antonio, ""Nativization of English words in Spanish using analogy" Proceeding of the 7th ISCA Speech Synthesis Workshop, Kyoto, Japan, September 2010

[5] Bonafonte, Antonio and Agüero, Pablo Daniel , "The UPC TTS Systems Description", Actas de las V Jornadas en Tecnologías del Habla, Bilbao, Spain,November 2008

[6] 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.

[7] 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.

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

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

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.