Adaptation of the URL-TTS system to the 2010 Albayzin Evaluation Campaign

Lluís Formiga, Alexandre Trilla, Francesc Alías, Ignasi Iriondo and Joan Claudi Socoró

GTM - Grup de Recerca en Tecnologies Mèdia.
La Salle - Universitat Ramon Llull. C/Quatre Camins 2, 08022 Barcelona, Spain
{llformiga,atrilla,falias,iriondo,jclaudi}@salle.url.edu

Abstract

This paper presents the text-to-speech (TTS) synthesis system of La Salle (Universitat Ramon Llull, URL) and its adaptation to the Albayzin Evaluation Campaign of FALA2010 conference. The URL-TTS system follows the classical scheme of unit selection TTS synthesis systems. However, it presents two distinguishable particularities: *i*) prosody prediction learned from labelled data by means of Case-Based-Reasoning (CBR) and perceptual weight tuning by means of active interactive Genetic Algorithms (aiGA). The aiGA-based weights are compared to multilinear regression (MLR) weights both considering classical averaged cost function and its root-mean squared variant. The internal validation tests and the results of the evaluation campaing are described, and finally discussed.

Index Terms: speech synthesis, unit selection, weight tuning, prosody prediction, interactive genetic algorithms, case-based reasoning

1. Introduction

The text-to-speech (TTS) synthesis system of the Grup de Recerca en Tecnologies Mèdia (GTM) of La Salle (Universitat Ramon Llull) (URL-TTS) is based on the original mid-90's second generation [1] Catalan concatenative TTS system, which considered diphones as basic units and *TD-PSOLA* for waveform generation [2, 3]. Subsequently, the system has been improved across years until the current unit selection TTS (US-TTS) synthesis system (see [4] for further details). The unit selection based URL-TTS synthesis engine presents two principal particularities (see figure 1): *i*) a case-based reasoning (CBR) prosody prediction module based on learning prosodic patterns from recorded corpora [5], and *ii*) a unit selection module, which integrates real human perceptual preferences through weights tuned by active interactive genetic algorithms (aiGA), which are adjusted at cluster level [6, 7, 8]. Moreover, great effort has been done to obtain automatic corpus development tools in order to speed up the set-up of the URL-TTS synthesis system for new voices [9]. This additional work involves features such as improving the selection of texts to be used during the recording process, including rules for avoiding ambiguity on phonetic transcription, refining unit segmentation [9] and reliable pitch marking [10]. In addition, there has been some further research focused on new acoustic parametrizations based on voice quality (VoQ) and harmonic plus noise models (HNM) [11], besides new approaches for expressive speech corpus parametrization [12]. All those improvements have been developed with the support of several research projects: SALERO, (IST-FP6-027122), SAVE (TEC2006-08043/TCM), evMIC (TSI-020301-2009-25).

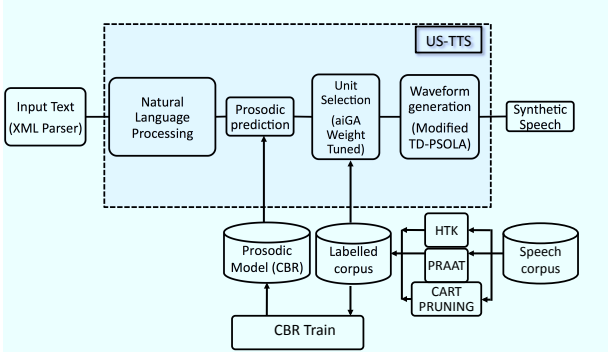


Figure 1: Block diagram of the URL-TTS synthesis system based on unit selection (US-TTS).

In this contest, we have incorporated some contributions with respect to the previous competition [4]: *i*), the speech corpus preparation includes a (quite simple) corpus pruning process based on detecting outlier voiced units with the aid of a of a clustering tool (the *wagon* tool of Festival [13]); *ii*) the prosody prediction module [5] has been incorporated in the TTS system, providing a richer prosodic reference for driving the unit selection process. This module has been used to guide the unit selection module only, and it has not been applied for conducting the posterior *TD-PSOLA*-based signal processing modification. In contrast, it has been considered the use of natural prosody of the retrieved acoustic units. This decision has been taken to recover the natural micro-prosody of units while minimizing the need of signal manipulation, following the classical idea of US-TTS systems to "*choose the best to modify the least*" [14]. Regarding the unit selection module, *iii*) the selection process has been updated through the use of 14 subcosts and 3 types of parameters (acoustic and linguistic parameters in target subcosts, and acoustic parameters in concatenation subcosts), and *iv*) the cost function weights are adjusted by using a three-stage process, which involves clustering and perceptual weight tuning [8]. Finally, *v*) TD-PSOLA is used for the waveform generation, minimizing both pitch and energy discontinuities around concatenation points.

This paper is organized as follows. Sections 2, 4, 5 and 6 describe the main modules of the current URL-TTS synthesis system based on unit selection. Section 3 is devoted to the Albayzin 2010 Evaluation Campaign corpus preparation process. The internal validation tests and the evaluation campaing results are presented in section 7, and conclusions and future work are outlined in sections 8 and 9.

2. Phonetic transcription

The *phoneme-set* used by the URL-TTS system is derived from SAMPA [15]. The phonetic transcription module consists of a rule-based system [16]. The rules are applied on a data structure that is a list of grapheme-phoneme pairs within a statement. It is possible to use insertion (I) or deletion (D) rules. The rules are applied only when the evaluation (E) of a phoneme characteristic yields a positive result.

$$E(gr ==' h') \quad \rightarrow \quad D(gr) \tag{1}$$
$$E(gr ==' x') \quad \rightarrow \quad I(/ks/) \tag{2}$$

Rule (1) indicates that the grapheme (gr) $'h'$ must be deleted, while rule (2) indicates that the grapheme $'x'$ must be transformed into the phonemes pair $/ks/$, thus implying a phonetic insertion. Regarding the exceptions, the system includes a dictionary that is consulted before applying any rule.

3. Creation of Unit Inventory

The voice of this competition (internally named as uvig_da_es) has been the 4th Spanish voice adapted to the URL-TTS synthesis system. Previously, we created the url_sam_es voice for weather forecasting restricted domain, the url_pat_es emotionated voice with 5 basic emotions (anger, joy, neutral, sad and sensual), and adapted the upc_ma_es voice for the 2008 Albayzin competition. It has to be added that url_pat_es voice is the main voice for the Spanish version of the URL-TTS synthesis system, which has been used in the projects mentioned in the introduction. In addition, it is worth noting that URL-TTS system is multilingual. The system also supports 2 Catalan voices (upc_pau_ca and upc_ona_ca from FesCat) and 4 English voices (url_sam_en, url_pgp_en, url_lau_en and url_rog_en) making a total of 8 public voices available. All of them may be tested on the GTM public website[1].

3.1. Segmentation and Labelling

3.1.1. Phonetic segmentation with pauses detection

In the segmentation process, the speech corpus is labelled indicating the temporal limits at the phoneme level. Our research group has been working to improve the segmentation process in recent years, in terms of the quality of the labelling process, the ease of use with the inclusion of user interfaces and language independence. Presently, the training and the posterior segmentation processes are based on Hidden Markov Models (HMM). To this end, a proprietary *Matlab*® code has been developed, also using the HTK tool (Hidden Markov Model Toolkit) [17].

The corpus that has been provided for the 2010 Albayzin Evaluation Campaign has been recorded by a male in neutral voice. It consists of 1217 utterances, 17797 word instances and a total vocabulary size of 5465 words. Regarding its analysis for the competition, the apparition and omission of silences have been controlled. Therefore, the pauses are correctly set according to the text. Alternatively, occlusive sounds are treated in special so that voice bursts and the previous silences are modeled as different units. At the end of the process we have obtained 826 different diphones with a total of 88571 diphones on the corpus.

[1] http://www.salle.url.edu/portal/departaments/home-depts-DTM-projectes-demos

3.1.2. Pitch marking

The PRAAT tool [18] has been used for signal pitch marking. It performs an acoustic periodicity detection on the basis of an accurate autocorrelation method [19]. In a first step, the voiced parts of the spoken utterance are pitch marked using this procedure. The pitch mark values are allowed to range between 75 and 600 Hz. In a second step, the unvoiced parts of the spoken utterance are given a sequence of pitch marks corresponding to the linear interpolation between the values of the previous pitch mark and the following one.

3.2. Corpus *pruning*

The process of recording and automatically labeling (segmentation and pitch marking) a speech corpus is prone to make errors. During the recording process, the speaker may introduce variants in the pronunciation or changes in the speed of delivery. Hence, elisions may be performed by speeding-up the speaking rate, or breaks may be introduced in the case of slowing it down, among others. A low-rate error labeling process is crucial for the general success of our US-TTS synthesis system, since the unit selection process itself is not capable of guaranteeing the retrieval of an error free unit sequence. In contrast to considering an exhaustive manual revision, a quite simple pruning process that attempts to detect errors in the recording and labeling phases has been implemented. In this work, the pruning has been performed at the phoneme level, by only considering voiced phonemes as they present more consistent parameters for the analysis. For each phoneme, the pruning process takes into account its prosodic parameters (pitch, energy and duration) and the first 3 spectral formants (obtained with [18]). Next, the 6-dimensional space (3 prosodic dimensions plus 3 spectral dimensions) is clustered using the *wagon* tool of Festival [13]. Once the phoneme groups are defined, the labelled phonemes out of their corresponding region are removed. As a result, 4908 recorded units are removed from the overall 88571 units (i.e. a 5.54% corpus size reduction).

4. Prosody Prediction

The URL-TTS synthesis system incorporates a corpus-based method for the quantitative modelling of prosody [5], following the case-based reasoning (CBR) algorithm proposed by [20]. This module predicts three main prosodic parameters: the fundamental frequency (F0) contour, the segmental duration and the energy, with the purpose of guiding the unit selection.

The automatic extraction of prosodic features from text starts from our linguistic analysis tool [21]. It carries out the phonetic transcription of text (based on SAMPA), annotating intonation groups (IG), stress groups (SG), words and syllables. The IG in Spanish is defined as a structure of coherent intonation that does not include any major prosodic break [22]. Prosodic breaks take place due to pauses or significant inflections of the F0 contour. The SG is defined as a stressed word preceded by one or more unstressed words, if they appear.

For the F0 contour modelling, the SG has been chosen following the proposal of [23]. The SG incorporates the influence of the syllable (it includes one stressed syllable plus some unstressed ones) and the pitch structure at IG level is achieved by the concatenation of SG contours. However, this model lacks variations due to micro-intonation. Up to now, we only differentiate between declarative, exclamatory, interrogative and suspended/unfinished IGs [24], which can be reliably identified from punctuation signs. Another attribute is the placement of

the tonic syllable in the SG. Finally, other considered attributes are the number of syllables of the SG and the positions of the SG relative to the IG and the sentence.

A quantitative representation of the F0 contour has been used, by means of the coefficients of the polynomial that minimizes the error between the original set of points and the polynomial. Therefore, F0 parameters consist of the coefficients of the polynomial that are adjusted to minimize the distance between the polynomial and a collection of points that represent the value of the average F0 of every phoneme. This mean value of F0 is referenced to the centre of each phoneme of the IG.

For segmental duration and energy modelling, the phoneme has been chosen the basic acoustic unit (as [25, 26]). These parameters depend on basically the phoneme identity and the context where it is placed (attributes related to position and stress).

5. Unit Selection

5.1. Framework

The unit selection module follows the classical scheme described by Hunt and Black in [27]. The corpus units are retrieved by means of the Viterbi dynamic programming algorithm [28], which seeks the best sequence of units by minimizing a cost function. This cost function is defined as a weighted sum of several normalized subcosts (see equation (5)). In general terms, these subcosts are composed of target and concatenation measures [27]. For each possible candidate unit, target subcosts measure the difference between the ideal unit on that position (either by linguistic definition or prosodic prediction) and the candidate unit. Moreover, for each possible pair of candidate units, concatenation subcosts measure the acoustic discontinuity at the concatenation point.

Thus, the unit selection cost function of unit i jointly with unit j is defined by the following equations:

$$C_T(i) = \sum_{k=0}^{param\_t} w_T^k \cdot {SC_T}^k(i) \tag{3}$$

$$C_C(i,j) = \sum_{k=0}^{param\_c} w_C^k \cdot {SC_C}^k(i,j) \tag{4}$$

$$C(i,j) = C_T(i) + C_C(i,j) \tag{5}$$

where ${SC_T}^k(i)$ and ${SC_C}^k(i,j)$ represent target and concatenation subcosts, which are weighted by w_T^k and w_C^k, respectively, and they are computed as:

$${SC_T}^k(i) = D\left[P(u_i)^k, P(t_i)^k\right] \tag{6}$$

$${SC_C}^k(i,j) = D\left[P(u_i^R)^k, P(u_j^L)^k\right] \tag{7}$$

where u_i is the candidate unit, t_i is the target unit, u_i^R is the parametrization on the right concatenation point of the candidate unit and u_j^L is the parametrization of the left concatenation point of the candidate unit. $D[\cdot,\cdot]$ is the distance function (Manhattan, euclidean, cubic, etc.) and $P(\cdot)^k$ is the measured value of parameter k for the corresponding unit.

Moreover, for this particular competition, we wanted to analyse the effects of changing classical averaged cost function (AVG) (see equation (5)) [27] for the root mean squared (RMS) cost function variant proposed in [29]. RMS cost function considers quadratic weighted sum of different subcosts instead of computing the lineal weighted sum of subcosts (see equation (10)).

$$C_T(i) = \sum_{k=0}^{param\_t} \left(w_T^k \cdot {SC_T}^k(i)\right)^2 \tag{8}$$

$$C_C(i,j) = \sum_{k=0}^{param\_c} \left(w_C^k \cdot {SC_C}^k(i,j)\right)^2 \tag{9}$$

$$C(i,j) = \sqrt{C_T(i) + C_C(i,j)} \tag{10}$$

In terms of target subcosts, we consider four acoustic subcosts (pitch, energy and left/right half phone durations) and seven linguistic subcosts (position in utterance, position in word, position in syllable, previous and next phonemes, part-of-speech and syllable stress). That makes a total of 11 target subcosts. As concatenation subcosts, we consider discontinuity of pitch, energy and cepstral coefficients at the concatenation point. Cepstral distance is computed considering the first 12 Mel-Cepstral coefficients along their derivatives. Overall, the cost function is composed by 14 subcosts of 3 different types (acoustic and linguistic for target and acoustic for concatenation subcosts).

5.2. Weight Tuning

The weights w_T^k and w_C^k of the cost function are tuned by a 3-step process:

i) Automatic weight tuning is performed using Multilinear Regression (MLR) [27, 30]. In order to avoid negative weight values, we used non-negative least squares implementation [31]. For each recorded unit in the corpus, MLR performs regression across the 20 acoustically nearest units considering the cepstral distance and their related subcosts.

ii) Once unit weights are automatically tuned at unit level, in a second phase, these weights are clustered by expectation maximisation (EM) algorithm in order to obtain weight patterns for each cluster [32]. EM is chosen since it is the method that obtains better validation clustering indices [33]. Afterwards, phonetic and linguistic information of each unit is mapped to weight patterns clusters by means of a classification and regression tree (CART). At this point we have weight patterns at cluster level, where the cluster is defined by linguistic and phonetic specifications.

iii) In the final stage, the weights for each cluster are tuned perceptually. The number of clusters is set to 5 after reaching a consensus among different validity indices [34]. Once the groups of units are defined, four representative sentences of each cluster (mainly containing units of that cluster) are selected. The utterances are chosen through an entropy maximization algorithm [35]. These 20 sentences (4 sentences for each of the 5 clusters) are then used for conducting the perceptual weight tuning process based on active interactive Genetic Algorithm (aiGA), following the scheme described in [8]. It is worth noting that no prediction of prosody is considered for the weight tuning, assuming an ideal process by extracting the prosody values of the target sentence. Finally, the aiGA-based weights are obtained and a new CART tree is built for determining the final perceptual weights pattern per cluster.

6. Waveform Generation

The waveform generation process included in the URL-TTS synthesis system is based on *TD-PSOLA* [36]. In that original work, all units are pitch-synchronously resynthesized overlapping their frames in order to match the duration and pitch of the target unit sequence. Discontinuities of pitch are minimized by interpolating pitch marks around the concatenation points between units that are not consecutive in the corpus. In this work, informal listening tests have shown that the synthetic speech quality is better when the target F0 and duration are recovered from the corpus instead of considering the CBR-based prosodic prediction.The original pitch marks structure is kept in the speech segments generated from units that are consecutive in the corpus. At each concatenation point the signal frames are interpolated, following new pitch marks values in order to achieve a smoother pitch contour. Also, signal amplitude adjustment is conducted to avoid energy discontinuities.

7. Experiments

In this section, the experiments conducted to set-up the URL-TTS synthesis system and the 2010 Albayzin Evaluation Campaign results are described. The validation experiments are perceptual tests considering Mean Opinion Score (MOS) [37]. Some investigations [38, 39] state that pairwise direct comparison (pairwise preference tests) overcomes MOS in terms of obtaining preference for final users in the case of comparing similar systems. To that effect, we adapted the classical MOS methodology to a double stimuli input in order to obtain the advantages of both methods. That is, the same input utterance was presented to the user synthesized by two different TTS system configurations, but the user had to rank them independently instead of choosing which one was the best. For testing the stimuli, we used the TRUE platform [40], which is capable to perform MOS, pairwise comparison tests or both at the same time. After presenting the validation tests, the results collected from the evaluation campaing are described and discussed.

7.1. Validation of weights with copy-prosody

Once the weights have been perceptually tuned, they are submitted to a subjective validation process to confirm their appropriateness. We consider the weights obtained by MLR [30] as the baseline for validating the aiGA-based weights.

To that effect, 20 utterances different from the ones involved in the perceptual adjustment were chosen from the speech corpus to be part of a preference test. The utterances were synthesized by 4 different unit selection configurations (aiGA-rms, aiGA-avg, MLR-rms, MLR-avg). aiGA/MLR identifies the weights used and rms/avg identifies the cost function involved in the unit selection process. The original recorded prosody from the utterance (copy-prosody) was used, as done during the perceptual weight tuning stage. In addition, the units composing the utterance were removed from the corpus in order to avoid the selection of those units, and thus, obtain a more reliable evaluation of the compared unit selection processes. Moreover, the natural recorded version of the utterance was also presented to the evaluators along with each pair of stimuli in order to provide an ideal *target*.

Six evaluators participated in the validation tests, obtaining the results depicted in figure 2. As it can be observed, better synthesis is achieved by aiGA-based methods: their corresponding averaged MOS results are 3.54 for aiGA through AVG cost function and 3.41 through RMS cost function, although if

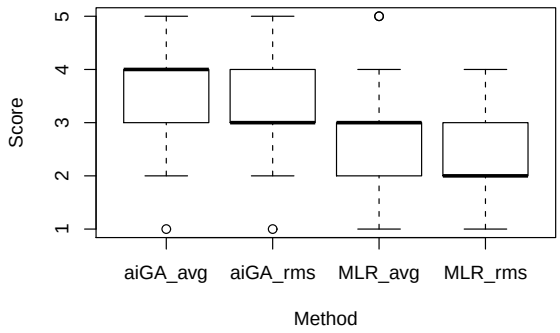


Figure 2: Internal MOS results comparing different weight tuning (aiGA / MLR) and two different integration cost functions (averaged vs. root-mean squared) when the target prosody is extracted from the recorded units.

the Bonferroni correction method is applied to test the significance of the results [41], we can conclude that their difference (0.13) is not statistically significant ($p = 0.743$). MLR-based weights behave significantly worse than the perceptual weights within both cost functions ($p < 0.001$). However, the AVG cost function computed with the MLR weights (MOS: 2.94) behaves slightly better than RMS cost function (MOS: 2.36) with their difference (0.58) being statistically significant ($p < 0.001$). As a last step, we also analyzed the pairwise comparison significance through a signed ranked test and we obtained the same results.

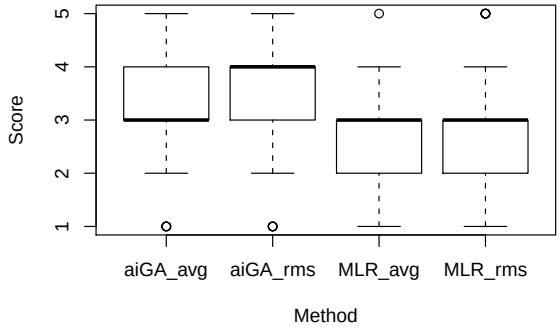


Figure 3: Internal MOS results comparing different weight tuning (aiGA / MLR) and two different integration cost functions (averaged vs. root-mean squared) when the target prosody is predicted by the CBR-based technique.

7.2. TTS final adjustment

In order to test the performance of the whole TTS synthesis system, we incorporated the CBR-based prosody prediction module with 20 utterances selected from the 2010 Albayzin Evaluation Campaign sets. As no natural prosody was available at that time, the natural recorded sentence was not presented to the evaluation users.

The same six evaluators participated in the final system validation tests, obtaining the new results depicted in figure 3. Again, better synthesis is achieved by aiGA-based methods: their corresponding averaged MOS values are 3.50 for aiGA through RMS cost function and 3.35 through AVG cost func-

Table 1: Groups detected by Bonferroni pairwise analysis

Group	Weight Tuning	Cost Function	Prosody	MOS Score
1	aiGA	AVG	COPY	3.54
	aiGA	RMS	CBR	3.50
	aiGA	RMS	COPY	3.41
	aiGA	AVG	CBR	3.35
2	MLR	AVG	COPY	2.94
3	MLR	RMS	CBR	2.61
	MLR	AVG	CBR	2.56
	MLR	RMS	COPY	2.36

tion, although their difference (0.15) is not statistically significant ($p = 0.517$). MLR-based weights again behave significantly worse than the aiGA-based weights within both cost functions ($p < 0.001$). However, in this case, the difference of averaged MOS values (0.05) between RMS (MOS: 2.56) and RMS cost functions (MOS: 2.61) is not statistically significant ($p < 1$).

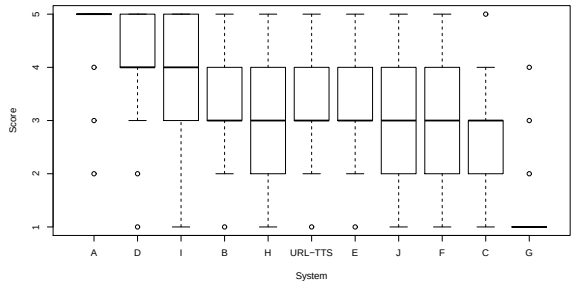
Next, the effects of including CBR-based prosody prediction to the unit selection module are discussed. The obtained results (copy-prosody and CBR-based prosody) were analyzed simultaneously. To that effect, we applied Bonferroni pairwise analysis in order to identify groups on the MOS evaluations. Groups are defined by configurations with no significant differences among them. The analysis found three groups in terms of the MOS results, as it can be seen on table 1. It can be observed that CBR-prosody prediction does not introduce major alteration to the copy-prosody results. Thus, the determinant factor for the URL-TTS synthesis system based on unit selection to obtain high quality speech is the weight tuning methodology. Under MLR-weight tuning methodology, synthesis with artificial (CBR-based) prosody is unable to reach the quality of natural prosody. Nevertheless, this difference is overcome by the aiGA-based weights.

As a result, the system presented to the 2010 Albayzin competition includes CBR prosody prediction, aiGA-based weight tuning and RMS cost function (as presented slight better results than AVG cost function, although not significant). This configuration achieved a MOS score of > 3.30 in the validation experiments.

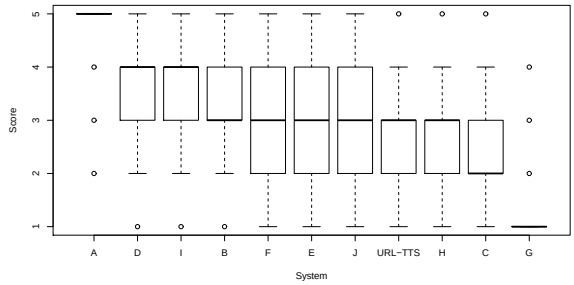
7.3. 2010 Albayzin Evaluation final results

Once the set-up of the system was completed, 400 synthesized sentences were presented to the 2010 Albayzin Evaluation Campaign. This evaluation campaign consisted of 3 separate analyses in order to assess different aspects of the evaluated TTS synthesis systems: *i*) similarity to the original recorded voice, *ii*) overall quality through mean opinion scores (MOS), and *iii*) intelligibility by computing word error rate (WER) on sentences composed of random words (i.e. with no clear meaning). The number of users involved in each test was substantially different depending on the test. Whether around 541 users were involved in the MOS test (see figure 4(b)), only around 137 users were involved in the voice similarity tests (see figure 4(a)) and 182 were involved in the WER tests.

In terms of similarity to the recorded voice, the URL-TTS system performs quite well since it is a US-TTS synthesis system, yielding a similar MOS value to the internal validation tests (average MOS$= 3.20$). However, on the overall quality MOS test, the URL-TTS decreases its score to 2.62. This significant decrease compared to the obtained MOS results may be motivated to several factors. Firstly, no natural voice was used on the validation tests, which makes the results not comparable. Secondly, few evaluators conducted the internal validation tests considering only relative improvements instead of considering the quality of other TTS systems. Finally, the URL-TTS synthesis system presents several intelligibility problems, reflected with a poor WER (0.31). This factor maybe caused by the presence of artifacts, that definitely affects the overall preceived synthetic speech quality. It is worth noting that, besides including a pruning process, the corpus creation has been fully automatic with no manual intervention at any stage of the process (neither using the given labelings or transcriptions).



(a) *Similarity to natural voice*



(b) *Overall quality*

Figure 4: FALA2010 results through different systems [42]

8. Conclusions

This paper describes the main advances included in the URL-TTS synthesis system with respect to the previous 2008 Albayzin competition. The two key elements are the CBR-based prosody model and the aiGA-based weight tuning. After several perceptual experiments, the URL-TTS synthesis system has obtained acceptable internal validation (MOS> 3.30) and similarity to the natural voice (MOS$= 3.20$) results. However, there has been a decrease on the overall quality according to the evaluation campaign results (MOS$= 2.64$), where the URL-TTS synthesis system has been challenged against to other TTS systems and some intelligibility problems (WER$= 0.31$). In favor of URL-TTS system, it is worth noting that these results were obtained after reasonable reduced time for the TTS set-up and tuning, thanks to the fully automatic voice building tools and

tuning platforms.

In terms of the weight tuning of the cost function, it can be concluded that weight tuning is one of the key factors in order to obtain good synthetic speech quality for the US-TTS synthesis system at hand. In addition, the results present a significant improvement when considering perceptual tuned weights (aiGA-based) with respect to using automatically trained weights (MLR-based). However, the substitution of the cost function from averaged to root-mean squared does not yield notable quality changes. Moreover, the perceptual results obtained after including the CBR-based predicted prosody during the TTS execution remain almost unaltered. However, it is worth noting that other key factors for obtaining high quality synthetic speech through US-TTS synthesis (e.g. segmentation and pitch marking, pruning methodology, waveform generation, etc.) have not been explicitly analyzed in this paper, leaving their analysis and optimization for future works.

9. Future work

Future work will be focused on improving the intelligibility and naturalness of the URL-TTS synthesis system, improving the corpus building tools and revising the database pruning process accordingly. In addition, this work will be focused on improving synthesis flexibility so as to modify the speech identity and expressiveness. In this regard, we are currently working on adapting an HNM (Harmonic-plus-Noise Model) library to the current US-TTS synthesis system. The main objectives are: *i*) considering the CBR prosody predictions, besides improving the quality of concatenations smoothing (avoiding artifacts), by fully exploiting the potentialities of the HNM through interpolation techniques, and *ii*), gradually improving the flexibility of the system (i.e. using speech conversion methods) but keeping the final synthesis similarity to natural voice as high as possible.

10. References

[1] P. Taylor, *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

[2] J. Camps, G. Bailly, and J. Martí, "Synthèse à partir du texte pour le catalan," in *Proc. 19èmes Journées d'Études sur la Parole*, Bruselas, Francia, 1992, pp. 329–333.

[3] R. Guaus, F. Gudayol, and J. Martí, "Conversión textovoz mediante síntesis PSOLA," in *Jornadas Nacionales de Acústica*, Barcelona, España, 1996, pp. 355–358.

[4] C. Monzo, L. Formiga, J. Adell, I. Iriondo, F. Alías, and J. Socoró, "Adaptación del CTH-URL para la competición Albayzin 2008," *V Jornadas en Tecnología del Habla*, pp. 87–90, 2008.

[5] I. Iriondo, J. C. Socoró, and F. Alías, "Prosody Modelling of Spanish for Expressive Speech Synthesis," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, Honolulu, EUA, Abril 2007, pp. 821–824.

[6] F. Alías, X. Llorà, L. Formiga, K. Sastry, and D. E. Goldberg, "Efficient interactive weight tuning for TTS synthesis: reducing user fatigue by improving user consistency," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. I, Toulouse, Francia, 2006, pp. 865–868.

[7] L. Formiga and F. Alías, "Extracting User Preferences by GTM for aiGA Weight Tuning in Unit Selection Text-to-Speech Synthesis," in *Computational and Ambiental Intelligence - Proceedings on 9th International Work-Conference on Artificial Neural Networks (IWANN)*. San Sebastián, Spain: Springer (LCNS), June 2007, pp. 654–661.

[8] L. Formiga, F. Alías, and X. Llorà, "Evolutionary process indicators for active IGAs applied to weight tuning in unit selection tts synthesis," in *IEEE Conference on Evolutionary Computation*. Barcelona, Spain: IEEE, July 2010, pp. 2322–2329.

[9] G. Kienast, G. Thallinger, R. Fach, S. M. Freixes, O. Mayor, T. Bürger, M. Yan, T. Stolt, R. Villa, M. Romeo, C. Goodman, M. Matthews, and L. Formiga, "Third annual on-line public report," http://www.salero.eu/media/pdf/del/SALERO-D10.5.6-PublicAnnualReport2008.pdf, Semantic AudiovisuaL Entretainment Reusable Objects, Deliverable 10.5.6, 2008.

[10] F. Alias and N. Munne, "Reliable Pitch Marking of Affective Speech at Peaks or Valleys Using Restricted Dynamic Programming," *Multimedia, IEEE Transactions on*, vol. 12, no. 6, pp. 481–489, 2010.

[11] C. Monzo, À. Calzada, I. Iriondo, and J. Socoró, "Expressive Speech Style Transformation: Voice Quality and Prosody Modification Using a Harmonic plus Noise Model," *Proceedings of Fifth International Conference on Speech Prosody, Chicago, USA*, 2010.

[12] I. Iriondo, S. Planet, J. Socoró, E. Martínez, F. Alías, and C. Monzo, "Automatic refinement of an expressive speech corpus assembling subjective perception and automatic classification," *Speech Communication*, vol. 51, no. 9, pp. 744–758, 2009.

[13] A. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *Proceedings of EuroSpeech*, Rhodes, Greece, 1997, pp. 601–604.

[14] M. Balestri, A. Paechiotti, S. Quazza, P. L. Salza, and S. Sandri, "Choose the best to modify the least: a new generation concatenative synthesis system," in *Proceedings of EuroSpeech*, vol. 5, Budapest, Hungary, 1999, pp. 2291–2294.

[15] J. C. Wells, *SAMPA computer readable phonetic alphabetHandbook of Standards and Resources for Spoken Language Systems*. Berlin and New York: Mouton de Gruyter, 1997, ch. SAMPA computer readable phonetic alphabet, pp. Part IV, section B.

[16] I. Iriondo, "Producción de un corpus oral y modelado prosódico para la síntesis del habla expresiva," Ph.D. dissertation, Universitat Ramón Llull, 2008.

[17] "HTK," in *Recuperado el 19 de 09 de 2008, de http://htk.eng.cam.ac.uk*, 2008, pp. 149–150.

[18] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1.37)," http://www.fon.hum.uva.nl/praat/, 2010, as of July 1, 2010.

[19] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the Institute of Phonetic Sciences 17*, Amsterdam, The Netherlands, 1993, pp. 97–110.

[20] A. Aamodt and E. Plaza, "Case-based reasoning: foundational issues, methodological variations, and system approaches," *Artificial Intelligence Communications*, vol. 7, no. 1, pp. 39–59, 1994.

[21] S. Sánchez, "Sincat/2. lenguaje para la conversión grafema-fonema," Enginyeria i Arquitectura La Salle, Tech. Rep., 1997.

[22] J. M. Garrido, "Modelling spanish intonation for text-to-speech applications," Ph.D. dissertation, Departament de Filologia Espanyola. Facultat de Lletres. Universitat Autònoma de Barcelona, 1996.

[23] D. Escudero and V. Cardeñoso, "Applying data mining techniques to corpus based prosodic modeling," *Speech Communication*, vol. 49, no. 3, pp. 213–229, March 2007.

[24] F. Campillo and E. Rodríguez, "A method for combining intonation modelling and speech unit selection in corpus-based speech synthesis systems," *Speech Communication*, vol. 48, no. 8, pp. 941–956, 2006.

[25] E. Navas, I. Hernáez, and J. M. Sánchez, "Modelo de duración para conversión texto a voz en euskera," *Procesamiento del Lenguaje Natural*, vol. 29, pp. 147–152, 2002.

[26] J. P. Teixeira and D. Freitas, "Evaluation of a segmental durations model for tts," in *Computational Processing of the Portuguese Language: 6th International Workshop, PROPOR 2003, Faro, Portugal, June 26-27, 2003. Proceedings*, ser. Lecture Notes in Computer Science, N. Mamede, J. Baptista, I. Trancoso, and M. Nunes, Eds., vol. 2721. Heidelberg: Springer, 2003, pp. 40–48.

[27] A. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, Atlanta, EUA, 1996, pp. 373–376.

[28] A. Viterbi, "Error bounds for convolutional codes and an asymtotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, pp. 260–267, 1967.

[29] T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano, "An evaluation of cost functions sensitively capturing local degradation of naturalness for segment selection in concatenative speech synthesis," in *Speech Communication, Elsevier*, vol. 48, no. 247, 2006, pp. 45–56.

[30] Y. Meron and K. Hirose, "Efficient weight training for selection based synthesis," in *Proceedings of EuroSpeech*, vol. 5, Budapest, Hungary, 1999, pp. 2319–2322.

[31] C. Lawson and R. Hanson, *Solving least squares problems*. Society for Industrial Mathematics, 1995.

[32] C. Fraley and A. Raftery, "MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering," 2006.

[33] S. Günter and H. Bunke, "Validation indices for graph clustering," *Pattern Recognition Letters*, vol. 24, no. 8, pp. 1107–1113, 2003.

[34] K. Kryszczuk and P. Hurley, "Estimation of the Number of Clusters Using Multiple Clustering Validity Indices," *Multiple Classifier Systems LNCS*, pp. 114–123, 2010.

[35] J. Gauvain, L. Lamel, and M. Eskénazi, "Design Considerations and Text Selection for BREF, a large French read-speech corpus," in *First International Conference on Spoken Language Processing*. Citeseer, 1990.

[36] E. Moulines and F. Charpentier, "Pitch-Sinchronous waveform processing techniques for text-to-speech synthesis using diphones," in *Speech Communication*, vol. 9, 1990, pp. 453–467.

[37] ITU-T, "Methods for subjective determination of transmission quality," Recommendation ITU-T P.800, Geneve, Suïssa, 1996.

[38] Y. Alvarez and M. Huckvale, "The reliability of the ITU-T P. 85 standard for the evaluation of text-to-speech systems," in *Seventh International Conference on Spoken Language Processing*. Citeseer, 2002.

[39] D. Sityaev, K. Knill, and T. Burrows, "Comparison of the ITU-T P. 85 Standard to Other Methods for the Evaluation of Text-to-Speech Systems," in *Ninth International Conference on Spoken Language Processing*, 2006.

[40] S. Planet, I. Iriondo, E. Martínez, and J. Montero, "TRUE: an online testing platform for multimedia evaluation," in *Programme of the Workshop on Corpora for Research on Emotion and Affect*. Citeseer, 2008, p. 61.

[41] Y. Hochberg, "A sharper Bonferroni procedure for multiple tests of significance," *Biometrika*, vol. 75, no. 4, p. 800, 1988.

[42] F. Méndez Pazó, L. Docío-Fernández, M. Arza Rodríguez, and F. Campillo Díaz, "The Albayzín 2010 Text-to Speech Evaluation," in *Proceedings of Fala2010*, 2010.