# Aholab Speech Synthesizers for Albayzin2010

*I. Sainz, D. Erro, E. Navas, I. Hernáez, J. Sánchez, I. Saratxaga, I. Odriozola, I. Luengo*

Aholab Signal Processing Laboratory, University of the Basque Country, Spain
{inaki, derro, eva, inma, ion, ibon, igor, ikerl}@aholab.ehu.es

## Abstract

This paper describes the two Text-to-Speech (TTS) systems presented by Aholab-EHU/UPV in the Albayzin2010 evaluation campaign. The first system is a statistical parametric TTS based on HTS, with the incentive of using our own vocoder. The other one is a hybrid system in which we try to take advantage of the consistency of the statistical averaging and the segmental naturalness of the unit selection approach. It uses the acoustic parameters generated by the statistical system as the target sequence during the unit selection process. Informal listening tests and some objective measures show that adding the Intonation Break information during the voice building process improves the performance of both systems.

**Index Terms**: speech synthesis, statistical parametric, unit selection, evaluation

## 1. Introduction

The Albayzin TTS evaluation compares the performance of different TTS systems built with a common Spanish speech database. This year is the second edition of Albayzin, as well as our second participation in it. After seven weeks for voice building, participants are asked to synthesize several hundred test texts that will be evaluated to determine the quality of the synthetic voices in terms of: naturalness, similarity to the original speaker and intelligibility.

*AhoTTS* [1] is the synthesis platform for commercial and research purposes that Aholab Laboratory has been developing since 1995. It has a modular architecture, and written in C/C++ it is fully functional in both UNIX and Windows operating systems. Up to this date, synthetic voices for Basque, Spanish (Albayzin2008 voice) and English languages have been created.

This paper is organized as follows. First, we describe the two systems presented. In Section 3 the voice building process is explained. The evaluation results are presented and discussed in Section 4. And finally, some conclusions are drawn in Section 5.

## 2. Systems Overview

In order to take part in Albayzin 2010 evaluation, we have developed two TTS systems: an HMM-based system and a hybrid one [2] [3] [4]. Both systems share the linguistic analysis module. Besides, the parametric output of the statistical system is used as an input of the hybrid system. Therefore, instead of explaining each system on its own, a sequential analysis of the hybrid TTS synthesis process is going to be described in this section.

The architecture of the hybrid is shown in Figure 1. In short, HTS [5] output is used as target prediction in the unit selection module. Pitch and duration predictions from HTS are combined with internal ones and spectrum parameters are used in order to calculate the distance between target and candidate units. Our hybrid approach tries to combine the robustness of the average modelling with the segmental quality of natural speech units.

### 2.1. Language Processing

This first module performs several language dependent tasks. Text normalization and grapheme to phoneme conversion are conducted by means of rules, whereas POS tagging uses a specific lexicon and some simple disambiguation rules.

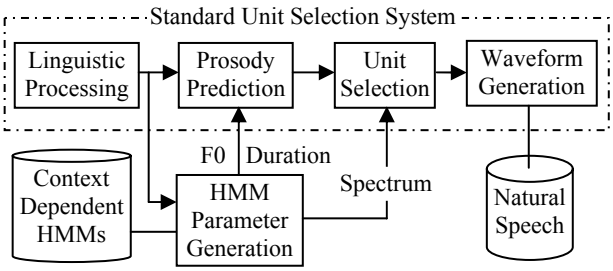


Figure 1: *Hybrid TTS Architecture*

### 2.2. Speaker-dependent HTS

Aholab had already built an HMM-based TTS system for Basque using HTS [6]. As HTS does not perform any kind of linguistic analysis, the output of the first module of AhoTTS had to be translated into proper labels containing phonetic and linguistic information. Taking into account that Basque language includes all the Spanish phonemes, only minor changes were necessary in order to adapt that system to Spanish (including the incorporation of Intonation Break feature (IB, see section 2.3.1)). The following features have been encoded into the context labels used by HTS:

- **Phoneme level:**
  - SAMPA label of the current phoneme.
  - Labels of 2 phonemes to the right and 2 phonemes to the left.
  - Position of the current phoneme in the current syllable (from the beginning and from the end).
  - Position of the current phoneme after the previous pause and before the next pause.
  - Position of the current phone after the previous IB and before the next IB.
- **Syllable level**:
  - Number of phonemes in current, previous and next syllables.
  - Accent in current, previous and next syllables.
  - Stress in current, previous and next syllables.
  - Position of the current syllable in the current word (from the beginning and from the end).
  - Position of the current syllable in the current accent group.
  - Position of the current syllable in the current sentence.
  - Position of the current syllable after the previous pause and before the next pause.

  - Position of the current syllable after the previous IB and before the next IB.
- **Word level:**
  - Simplified part-of-speech tag of the current, previous and next words (content/function).
  - Number of syllables of the current, previous and next words.
  - Position of the current word in the sentence (from the beginning and from the end).
  - Position of the current word after the previous pause and before the next pause.
  - Position of the current word after the previous IB and before the next IB.
- **Accent level:**
  - Type of current, previous and next accent groups, according to the accent position.
  - Number of syllables in current, previous and next accent groups.
  - Position of the current accent group in the sentence (from the beginning and from the end).
  - Position of the current accent group after the previous pause and before the next pause.
- **Pause context level:**
  - Type of previous and next pauses.
  - Number of pauses to the right and to the left.
- **Sentence level:**
  - Type of sentence.
  - Number of phonemes.
  - Number of syllables.
  - Number of words.
  - Number of accent groups.
  - Number of pauses.

In order to extract the framewise parametric representation of both the spectrum and the excitation, an HNM (Harmonics plus Noise Model) is used [7] that allows the reconstruction of speech too.

## 2.3. Prosody Prediction

This module performs several sequential tasks as IB insertion, duration prediction and intonation modelling. We have decided not to use our phrasing algorithm because its performance is still poor (too many false insertions that spoil the synthesis output). Therefore, we rely only on orthographic marks to assign phrase breaks. Nevertheless, thanks to the new IB prediction module, the absence of pause breaks in long word sequences is somehow alleviated. Besides, being the IB a more subtle phenomenon than the phrase break is, its miss-insertions are also less disturbing.

### *2.3.1. Intonation Break*

IB is an important phenomenon not only related to the intonation contour (e.g. F0 reset), but also to the duration (e.g. syllable lengthening) and acoustic realization (e.g. relaxed pronunciation) of phonemes adjacent to this event. As the corpus provided by the Albayzin organization included IB labels, we have built a CART that predicts their location from input plain text [8]. Among the features used to accomplish that goal, the following ones can be highlighted: POS in a three word window around current word, and the number of syllables, words and accent groups to previous and next breaks (IB or pause). The IB information is used in both prosody prediction and unit selection acoustic module at several unit levels (phoneme, syllable and word).

### *2.3.2. Corpus Based Intonation*

Our unit selection intonation modelling uses the voiced phoneme as the basic unit in a similar approach to [9]. Such a small unit provides greater flexibility, although the concatenations of non consecutive units inside syllables are significantly restricted. We implement a generic Viterbi search to find the sequence of candidate units from the database that minimizes a function cost composed by the target and concatenation subcosts [10] as shown in the following equations:

$$C(t,u) = \alpha \sum_{i=1}^{n} C^T(t_i,u_i) + (1-\alpha) \sum_{i=1}^{n-1} C^C(u_i,u_{i+1}) \quad (1)$$

$$C^T(t_i,u_i) = \sum_{j=1}^{P} w_j^T C_j^T(t_i,u_i) \quad (2)$$

$$C^C(u_i,u_{i+1}) = \sum_{j=1}^{Q} w_j^C C_j^C(u_i,u_{i+1}) \quad (3)$$

Where $t_i$ are target units and $u_i$ candidate ones. $C^T$ and $C^C$ are the target and concatenation cost respectively; $w_j$ is the j-th weight of the P target subcosts and the Q join subcosts. The main features employed in the target function are these: Type of proposition, Type of Accent Group (AG), Segmental characteristics of neighbouring phonemes, Position (in the AGs, syllable, word and phonic group), Accent, Duration, IB boundary.

Target weights are adjusted using a similar approach to the one proposed in [10] for acoustic unit selection. We first measure the pitch distance between units in the database and the relative distance regarding the adjacent voiced units. Then, we try to predict that distance as the summation of the target subcosts defined above, solving the weights as a multiple linear regression problem.

When two intonation units are not consecutive in the corpus, the following join subcosts are calculated: Pitch difference at the join, Pitch difference among natural neighbours of the units to be concatenated.

Join weights are manually assigned and some penalizations are added in order to hinder the concatenation of non consecutive voiced units inside a syllable, and to a lesser extent, inside an AG. Finally, the intonation contour is combined with the one predicted by the MSD-HSMM (multi-space distribution hidden semi-Markov models) modelling output from HTS. We just perform a weighted linear combination of both pitch contours, after phone alignment and interpolation in unvoiced regions. That way, we try to take advantage of the consistency of the statistical averaging and the segmental naturalness of the unit selection approach.

### *2.3.3. Duration*

CART zscore duration models were trained for voiced and unvoiced consonants, whereas Random Forests [11] were preferred for the vowels. In both cases, the same features were used: phoneme characteristics in a five phoneme window, stress, position (in syllable, word, IB and sentence), simplified POS, etc. Once again, the durations are combined with the ones predicted by HTS.

## 2.4. Acoustic Engine

Our acoustic engine performs the usual steps in a corpus-based concatenative system: pre-selection of candidate units, a dynamic programming step combining weighted join and target costs, and a concatenation step joining the selected units to form an output speech waveform. Halfphones are

selected as the basic unit because of the flexibility they provide to form longer units.

In our hybrid approach, the spectral parameters generated by the statistical parametric synthesis are used as the target during the unit selection process, combined with prosody and linguistic features.

#### *2.4.1. Unit Selection*

Target cost function (2) is divided in various subcosts which are calculated at the halfphone level: Phoneme context, Pitch and its slope, Duration, Accent, Type .of proposition, Position. A new subcost is added for the Hybrid System:

- *Spectral Distance*: Frame based Euclidean distance between target (HTS output) and candidate units after DTW [12] alignment. The distance is manually weighted according to three reduced phonetic classes: vowels, voiced and unvoiced consonants.

The concatenation cost function (3) is composed of seven subcosts, all but the *inter-syllable pitch range* being only computed for non-consecutive units: Pitch, Inter-syllable pitch range, Duration, Power, Spectrum, Voiceness, and Penalizations depending on the transition type.

Target weights are adjusted solving a multiple linear regression problem, as stated previously for the pitch modelling. The Euclidean distance of MFCC parameters is used as the predictee and the subcosts as the predictors. Different weights are estimated for left and right halfphones and for each phoneme type. Concatenation weights and α from equation (1) are adjusted manually.

#### *2.4.2. Waveform Generation*

The selected candidate units are joined using glottal closure instant information to get smooth concatenations. It is well known that prosody modifications reduce the overall natural quality of the voice. Therefore, only minor prosody modifications are done by means of pitch synchronous overlap and add techniques. The energy is smoothed over non consecutive halfphone transitions and a gain contour is applied in order to normalize the amplitude in the middle of each phoneme.

## 3. Voice Building

Organizers provided a medium sized (two hours long) speech database [13] recorded at University of Vigo by a male voice talent. The database consists of 1217 phonetically balanced sentences, automatically extracted phone segmentation and IB labels.

The voice building process involves several sequential tasks that are performed almost automatically. After segmentation labels are ready, linguistic and acoustic features are extracted and then, unit selection databases and prosody models are built and weights are trained. The training process of the statistical parametric voice is automatically done, once proper questions to build the trees are set.

### 3.1. Segmentation

Although the organizers provided segmentation labels, we decided to segment the whole corpus again with HTK toolkit [14]. Before doing so, transcriptions of some foreign words were manually corrected (e.g. West Side Story, pronounced by the speaker as B.w.e.s.t-s.a.j.T-e.s.t.o.r.i). Then, tied-state triphone models were trained and new labels obtained by means of forced alignment. Finally, pause boundaries were automatically refined with a simple processing based on phone duration and energy threshold. No manual revision of the segmentation labels was done.

### 3.2. Feature Extraction

All the language related features were extracted from our linguistic processing module. The extraction of the acoustic features consists of several steps. First, power normalization is performed by measuring the mean power in the middle of the vowels for all the sentences, and then normalizing each inter-pause interval. Then, pitch contour is detected combining three different methods in order to avoid gross errors (our own PDA (Pitch Detection Algorithm) [15], get_f0 from Snack Toolkit and Praat). HTK is used to generate 13 MFCC parameters calculated with a fixed 5ms frame. As far as the HTS training is concerned, the following parameters are extracted: f0 + 40 MFCCs.

### 3.3. Impact of IB information

As we had never used IB information during the voice building, a shallow analysis of its impact has been done. Questions related to IB appear in the upper third of all trees trained with HTS: duration, logF0 and MGC (Mel Generalized Cepstrum). The same can be said for the Hybrid TTS: the correlation of the duration prediction improves and IB information is taken into account in the trained weights for prosody and acoustic module (especially at phoneme level and for vowels). Informal listening also revealed that including the IB information had a positive impact in the performance of both systems.

## 4. Evaluation Results

Each listener completed three evaluation tasks: (i) Mean Opinion Score (MOS) to measure the similarity with the original voice, (ii) naturalness MOS, (iii) and an intelligibility test in which evaluators were asked to transcribe the Semantically Unpredictable Sentences (SUS) they heard. Up to 132 listeners completed the whole test. These were their main characteristics: 44% were experts on speech technologies, 92.5% were native speakers and 82% used headphones. Ten synthetic systems took part in the evaluation (identified with letters B-G). Natural voice (letter A) was also evaluated in order to fix the ceiling score.

In the present section detailed results are shown for our two systems in each of the three evaluation tasks. Unless the contrary is expressed, results from all listeners are analyzed. System ranking or grouping is based on the pairwise Wilcoxon test provided by the organization [16], which is a useful tool to know whether differences among systems are statistically significant or not.

### 4.1. Similarity Test

It measures the similarity to the original voice in a likert type scale ranging from 1 (*Sounds like a totally different person*) to 5 (*Sounds like exactly the same person*). The results for all the listeners are shown in Figure 2, whereas results for different groups of evaluators are displayed in Figure 3.

#### *4.1.1. Hybrid-TTS*

Our Hybrid system (letter D) obtained the best results (4.07 MOS) together with system I. That ranking remains constant for the different listeners' characteristics, although the expert evaluators scored our system a litter higher than the non-experts (4.19 and 3.99 respectively). Being our system a concatenative one (i.e. concatenates segments of natural speech) this high similarity to the original speaker could be

expected. However, listeners tend to score not only the segmental similarity but the supra-segmental one (prosody), and concatenation artifacts may play an important role in the subjective evaluation too. In any case, there is still a significant performance gap with respect to the system A (4.83 MOS).

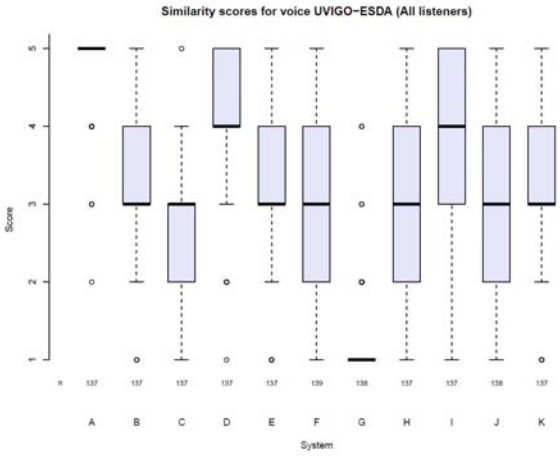


Figure 2: *Similarity to the original voice, all listeners*.

### *4.1.2. HTS-based system*

Our system (letter F) gets a MOS of 2.91, and it shares not significant differences with a group of 5 systems (H, K, E, J and C). It seems that the vocoding nature of the system has slightly degraded the similarity to the original voice. Nevertheless, we think that this section of the test is the least important one for typical TTS applications. Besides, it must be stated that the scoring is almost the same for both expert and non-expert evaluators (and the same occurs in the second evaluation task).

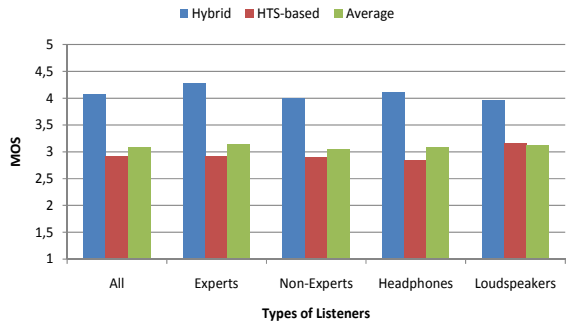


Figure 3: *Similarity to the original voice, different groups*.

## 4.2. Naturalness Test

It measures the naturalness of the systems in a likert type scale ranging from 1 (*Completely Unnatural*) to 5 (*Completely Natural*). The results are displayed in Figures 4 and 5.

### *4.2.1. Hybrid TTS*

Our Hybrid TTS is significantly more natural than the rest of synthetics systems, with a 3.71 MOS. Once again, the scoring of expert evaluators (3.89) is higher than the one from non-experts (3.69). And the gap is even larger between listeners that used headphones (3.86 MOS) or loudspeakers (3.4 MOS). We think that the hybrid approach has succeeded in improving the consistency that unit selection systems usually lack. Just one bad join or incorrectly labeled unit can spoil a whole sentence. Introducing the spectral output of the HMM-based system in the unit selection algorithm has alleviated that problem. Besides, combining two prediction methods produced a more robust prosody.

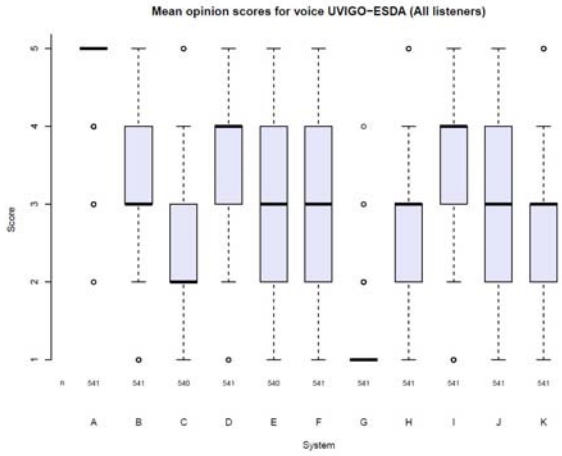


Figure 4: *Naturalness, all listeners*.

### *4.2.2. HTS-based system*

Our statistical system obtained a MOS of 3.15, ranking as the fourth best TTS in this task, together with system E. The robustness of the statistical averaging in the modelling process has yielded quite good results. And the same can be said as far as the performance of our own vocoder is concerned.

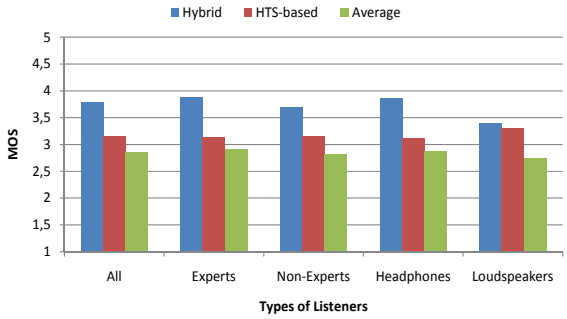


Figure 5: *Naturalness, different groups*.

## 4.3. Intelligibility Test

The organizers computed Word Error Rates (WER) for SUS as a measure of intelligibility. Unfortunately, no natural speech stimuli were available during the test due to the special structure of sentences needed. So it was not possible to make a comparison between synthetic and natural speech.

Non-native listeners might have introduced some noise in the evaluation (i.e. word errors due to their insufficient knowledge of the language). Therefore, they were not taken into account in the results presented here and displayed in Figure 6. Figure 7 shows WER for different listeners' groups.

### *4.3.1. Hybrid TTS*

It achieved a WER of 17%, obtaining the best results together with systems E, B, J and F. The Hybrid approach seems to have alleviated the problems caused by labeling errors or poorly pronounced units, yielding a good performance.

### *4.3.2. HTS-based system*

It managed a WER of 16%, obtaining the best results together with systems E, B, J and D. As happened in Albazyn2008

evaluation campaign, statistical modelling has yielded a pretty robust performance.

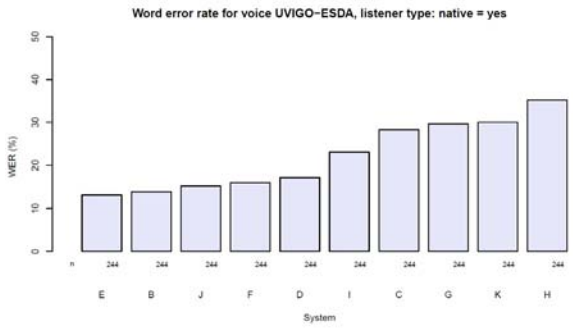


Figure 6: *WER, native listeners*.

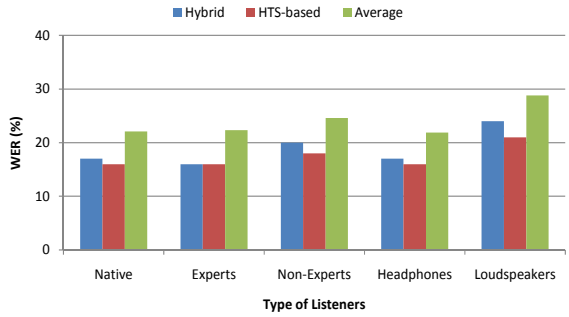


Figure 7: *WER, different groups*.

## 5. Conclusions

This has been our second participation in the Albayzin TTS evaluation campaign. Two synthetic voices have been built this year. One the one hand, an HTS-based TTS with a vocoder based on a parametric representation extracted from an HNM analysis. On the other hand, a Hybrid system that tries to combine the strong points of statistical and unit selection synthesis (i.e. robustness and segmental naturalness respectively). During the voice building process we introduced a feature we had never used so far: IB. We believe that its inclusion had a positive effect in the performance of both TTSs.

The Hybrid system got the best results (alone or together with other systems) in all the three sections of the evaluation, for experts and non-expert listeners. Those were very promising results, being this our first attempt to build a hybrid TTS. The HTS-based system scored pretty well too, above average in all sections but the first one.

A considerable gap between natural and synthetic voices still exists, but hybrid approaches seem to be an appropriate way to try to make the margin smaller.

## 6. Acknowledgements

The authors would like to thank the organizers of Albayzin TTS 2010 and the developers of all the tools employed during the voice building process.

This work has been partially supported by UPV/EHU (Ayuda de Especialización de Doctores), the Spanish Ministry of Science and Innovation (Buceador Project, TEC2009-14094-C04-02) and The Basque Government (Berbatek, IE09-262).

## 7. References

[1] I. Hernáez, E. Navas, J. Murugarren, and B. Etxebarria, "Description of the AhoTTS System for the Basque Language," *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*, ISCA, 2001.

[2] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda, "XIMERA: A new TTS from ATR based on corpus-based technologies," *Fifth ISCA Workshop on Speech Synthesis*, r, 2004, pp. 179-184.

[3] S. Rouibia, O. Rosec, and T. Moudenc, "Unit Selection for Speech Synthesis Based on Acoustic Criteria," *Text, Speech and Dialogue*, 2005, pp. 281-287.

[4] T. Hirai and S. Tenpaku, "Using 5 ms segments in concatenative speech synthesis," *Fifth ISCA Workshop on Speech Synthesis*, 2004, pp. 37-42.

[5] [Online], "HMM-based Speech Synthesis System (HTS)", http://hts.sp.nitech.ac.jp/

[6] D. Erro, I. Sainz, I. Luengo, I. Odriozola, J. Sánchez, I. Saratxaga, E. Navas, I. Hernáez, "HMM-based Speech Synthesis in Basque Language using HTS," *Proceedings of Fala2010*, Vigo: 2010.

[7] D. Erro, I. Sainz, I. Saratxaga, E. Navas, I. Hernáez "MFCC+F0 Extraction and Waveform Reconstruction using HNM: Preliminary Results in an HMM-based Synthesizer," *Proc. De las VI Jornadas en Tecnología del Habla*, Vigo: 2010.

[8] F. Campillo, J. van Santen, and E. Banga, "Integrating phrasing and intonation modelling using syntactic and morphosyntactic information," *Speech Communication*, vol. 51, 2009, pp. 452-465.

[9] A. Raux and A. Black, "A unit selection approach to f0 modelling and its application to emphasis," *ASRU, St Thomas, US Virgin Islands*, 2003.

[10] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *ICASSP '96: Proceedings of the Acoustics*, vol. 1, 1996, pp. 373-376.

[11] L. Breiman, "Random forests," *Machine learning*, vol. 25, 2001, pp. 5-32.

[12] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, 1978, pp. 43-49.

[13] E.R. Banga, C. G. Mateo, "Documentation of the UVIGO_ESDA Spanish database", *Grupo de Tecnoloxías Multimedia*, Universidade de Vigo, Vigo, España, 2010.

[14] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland:"The HTK Book (for HTK version 3.2)", Cambridge University Engineering Department, 2002.

[15] I. Luengo, I. Saratxaga, E. Navas, I. Hernáez, J. Sanchez, and I. Sainz, "Evaluation of Pitch Detection Algorithms Under Real Conditions," *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, IEEE, 2007, pp. IV-1057-IV-1060.

[16] F. Méndez, L. Docío-Fernández, M. Arza, F. Campillo, "The Albayzín 2010 Text-to Speech Evaluation", *Proceedings of Fala2010*, 2010.