BUCEADOR hybrid TTS for Blizzard Challenge 2011

Iñaki Sainz[1], *Daniel Erro*[1], *Eva Navas*[1], *Jordi Adell*[2], *Antonio Bonafonte*[2]

[1] Aholab Signal Processing Lab - University of the Basque Country, Bilbao, Spain
[2] TALP Research Center - Universitat Politècnica de Catalunya, Barcelona, Spain
{inaki, derro, eva}@aholab.ehu.es, jordi.adell@upc.edu, antonio.bonafonte@upc.edu

Abstract

This paper describes the Text-to-Speech (TTS) systems presented by the Buceador Consortium in the Blizzard Challenge 2011 evaluation campaign. The main system is a concatenative hybrid one that tries to combine the strong points of both statistical and unit selection synthesis (i.e. robustness and segmental naturalness respectively). The hybrid system has reached results significantly above average as far as similarity and naturalness are concerned, with no significant differences with most of the systems in the intelligibility task. This clearly improves the performance achieved in previous participations, and shows the validity of the hybrid approach proposed. Besides, an HMM-based system was built for the ES1 intelligibility tasks, using an HNM-based vocoder.

Index Terms: speech synthesis, unit selection, statistical synthesis, hybrid TTS system

1. Introduction

The Blizzard Challenge is an evaluation that compares algorithm performance of different text-to-speech (TTS) systems built with a common speech database. After a few weeks for voice building, participants are asked to synthesize several hundred test utterances that will be evaluated with respect to naturalness, similarity to the original speaker and intelligibility.

Buceador Consortium is formed by two research groups with extensive experience in TTS: TALP from Universitat Politècnica de Catalunya (UPC), and Aholab from the University of the Basque Country. Ogmios [1] is the TTS of the former and AhoTTS [2] the one of the latter. Both groups have taken part separately in previous Blizzard Challenge campaigns [3][4].

In this joint effort, UPC has provided the linguistic processing module for English, whereas Aholab has developed the prosodic and acoustic modules for two different TTSs: a statistical parametric system and a hybrid one.

This paper is organized as follows. First, we describe the characteristics of the two systems. 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

Two systems were developed: a statistical parametric one based on HTS [5] and a *hybrid concatenative TTS* [6]. The former was used in the ES1 task and the latter as the main English voice (EH1). They both share the language processing module developed by UPC.

2.1. Language Processing

The main goal of the front-end of a TTS system is to transform the input text into explicit linguistic information which is used to select either the units (in unit selection back-ends) or the models (in statistical synthesis). In our system we have used Ogmios to do so. The system first tokenizes the text into standard words, numbers, acronyms, etc. and verbalizes them. The tokenizer segments the text into tokens by means of a set of rules (regular expressions) taking into account features such as white spaces, punctuation, case, and also with specific rules for dates, url, etc. The tokenizer also assigns to each token a label that identifies the type of token (e.g.: ordinal number, month, etc.). The next step, the verbalization, transforms tokens which are not found in the lexicon into standard words. The tokenization and verbalization rules cover the most frequent cases in English, as different types of numbers (ordinal, cardinal, currency, etc.), acronyms, dates, url, etc.

POS tagging is performed using a basic statistical tagger. The probability of POS tag sequences are derived using 1 million tokens from the Penn Treebank WSJ Corpus. For unknown words, CART is used to estimate the probability of each possible tag.

The pronunciation of each word is based on the Unisyn lexicon provided by the University of Edinburgh [7]. It consists of 110K word entries and it is coded to represent different dialects. In order to increase its coverage, the lexicon was extended with words which are present in the LC-STAR US-English dictionary [8] and are not included in Unisyn. This LC-STAR lexica is multilingual and includes, for each language, 50K common words and 50K names and other specific words. Even if both lexica (Unisyn and lc-star) represent the pronunciation using SAMPA, the transcription criterion is slightly different. As a consequence, only 30% of the words which are present in both lexica share the same phonetic transcription. In order to improve the compatibility between both lexica, a finite state transducer was inferred to transform the LC-STAR lexicon into the Unisyn lexicon. The process is basically the same that we apply to the grapheme-to-phoneme task.

First, the words which appear in both dictionaries are selected. The pronunciation of Unisyn lexicon is called the target pronunciation while the one from the LC-STAR is called source pronunciation. Both pronunciations are aligned by means of an EM algorithm to map each phoneme from LC-STAR to one from Unisyn. The probability of the bi-phoneme sequences is estimated using ngrams. Finally, given the source pronunciation, we choose the target pronunciation so that it maximizes the joint source-target probability. This is implemented using a non-deterministic finite state transducer. The use of this technique raised the compatibility of both dictionaries, evaluated in an independent set of words, to 83%. The pronunciation of unknown words is derived using a grapheme-to-phoneme finite-state transducer which is trained from the Unisyn lexicon using the same procedure: letters are the source information and phonemes are the target representation.

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, POS and phrase accent. In continuous speech the function words usually lose their accented form and the full vowels are reduced to the shorter vowels or schwa. Furthermore, a set of phonotactic hand-crafted rules was applied. These rules cover different phenomena from aspirated plosives, to consonant assimilation and elision.

In the training phase, pauses were introduced during the alignment process. The viterbi algorithm was constrained with the phonetic transcription, but an optional silence between words was allowed. In the operative phase the major breaks were predicted using our previous break/no-break decision tree classifier (CART). The classifier was trained on one of the TC-STAR baseline voices [9] which consist on 10 hours of speech read by a professional speaker.

2.2. Speaker-dependent HTS

Aholab had already built an HMM-based TTS system for Basque [10] and Spanish [11] based on HTS. The HNM-based vocoder presented in [12] is used to obtain the framewise parametric representation of the speech signals at three different levels: log-f0, Mel-cepstral coefficients, and maximum voiced frequency (MVF). This vocoder allows high-quality waveform reconstruction too. The test utterances were synthesized using *HTS_engine version 1.03,* modified in order to include the HNM-based vocoder.

2.3. Hybrid System

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

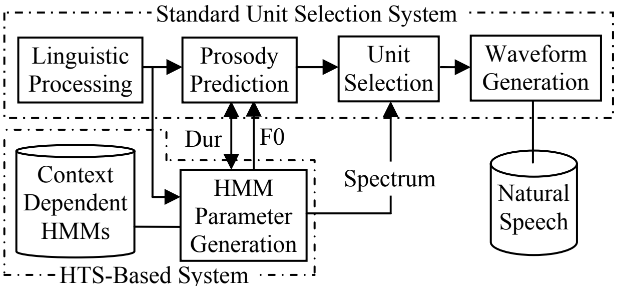

Figure 1: *Hybrid TTS Architecture*

2.3.1. Prosody Module

Most hybrid TTSs rely just on HMM's prosody prediction. However, better duration prediction can be achieved through the fusion of different techniques [13]. Besides, in [14] they got the best MOS (Mean Opinion Score) by imposing an external duration to the HMM-based intonation curve.

In our hybrid approach, a linear combination between HTS and CART duration predictions is performed. First, phone duration is predicted inside HTS engine. Then, this prediction is linearly combined with the one from the standard prosody module and forced at phone level. Finally, HTS predicts the length of each state inside the externally predetermined phone length. Three broad phoneme classes are taken into account during the fusion: voiced consonants, unvoiced consonants and vowels.

The fusion of the two intonation curves is performed in several stages. First, f0 values are linearly interpolated in unvoiced regions and both curves are segmented at phone level preserving only the f0 values of canonically voiced phonemes. For each voiced phoneme a 3 point pitch stylization is performed. Finally, a weighted linear combination is performed between aligned phone sized pitch portions.

This simple approach has yielded slight improvements in objective measures and statistically significant ones in subjective black box tests [6].

In the prosody fusion process explained above, weights of the linear combination had been manually tuned, giving more relevance to the HTS pitch prediction and to the CART duration prediction respectively, according to the results of objective tests done for other voices.

2.3.2. Unit Selection

During the unit selection process, most hybrid TTSs rely solely on the acoustic trajectories generated by the statistical parametric system. Contrary to that option, we maintain the usual linguistic and prosodic target sub-costs of the baseline system, adding just a new sub-cost:

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

The main advantage of this approach is that selecting the units by means of modeling both explicitly (output of HTS) and implicitly (linguistic target sub-costs) their acoustic similarity, seems a more robust procedure. One of the key contributions of the spectral distance is to prevent "bad units" (i.e. wrongly labeled or poorly pronounced) from being selected, achieving more consistent synthesis. As the computation of spectral distances is especially time-consuming, only linguistic and prosodic (and therefore much less complex) target sub-costs are used in a pre-selection stage, speeding up the synthesis process that way.

2.3.3. Waveform Generation

The selected candidate units are joined using glottal closure instant information so as to get smooth concatenations. It is well known that prosody modifications reduce the overall natural quality of the voice. So, having in mind the size of the corpus available, it was decided not to perform any kind of modification, but energy normalization.

3. Voice Building

This year, a large speech database was kindly supplied by Lessac Laboratories Inc for Blizzard Challenge 2011. The database consists of 12000 utterances recorded by a US female voice talent, lesseme labels and pitch marks. Lessemes are a symbolic representation that, apart from phonetic information, also include co-articulation and supra-segmental features to try to capture the musicality of speech [15]. In addition to this, prosodic breaks and operative words (that carry the highest pitch prominence) were marked in the available texts. But due to lack of time, none of that information was used during the voice building, although it would probably improve the quality of the synthesized speech.

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 database and prosody models are built and target weights are trained. The training process of the statistical parametric voice is automatically done, once proper questions for building the trees are set.

3.1. Segmentation

We decided not to use the lesseme format provided by the organizers. Therefore, standard US SAMPA phonetic labels were generated from plain text by Ogmios. A 16kHz sampling rate was used during the voice building process.

The whole corpus was segmented with HTK toolkit [16]. Tied-state triphone models were trained from a plain start, and phoneme labels were obtained by means of forced alignment. Finally, pause boundaries were automatically refined with a simple processing based on adjacent phone duration and energy threshold. The pruning of the database was performed on two levels: sentence and unit. First, sentences containing words with unknown POS, as well as a few hundred utterances with the worst alignment score were removed. Then, the models were retrained and a definitive segmentation was achieved. At the lower level, units were penalized during the selection process according to an outlier score based on: alignment score, spectral distance to the center of phonetic clusters and duration outliers. Both hybrid and statistical systems were built with these labels. No manual revision was done.

3.2. Feature Extraction

All the language related features were generated with the linguistic processing module of Ogmios. 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. Next, pitch contour is detected combining three different methods in order to avoid gross errors (Aholab's PDA (Pitch Detection Algorithm) [17], get_f0 [18] from Snack Toolkit and Praat [19]). 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 + MVF.

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, and (iii) an intelligibility test comprised of Semantically Unpredictable Sentences (SUS) and addresses. The first two tasks were divided into more subtasks according to the genre of texts to be synthesized: novel, news and reportorial.

12 synthetic systems took part in the evaluation (identified with letters B-M). B is the benchmark Festival unit selection system, C is the benchmark speaker-dependent HMM system and D is the same as C with 48kHz sample rate data. Natural voice (letter A) was also evaluated in order to fix the ceiling score.

In the present section detailed results for our system in each of the three evaluation tasks are shown. Unless the contrary is expressed results for all listeners are analyzed. The *average* system shown in the figures represents the mean score among all synthetic TTS. System ranking or grouping is based on the pairwise Wilcoxon test provided by the organization, which is a useful tool to know whether differences among systems are statistically significant or not. It was computed with a significance level of p=0,01 and Bonferroni correction.

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 and different types of input texts are shown in Figure 2 and Figure 3.

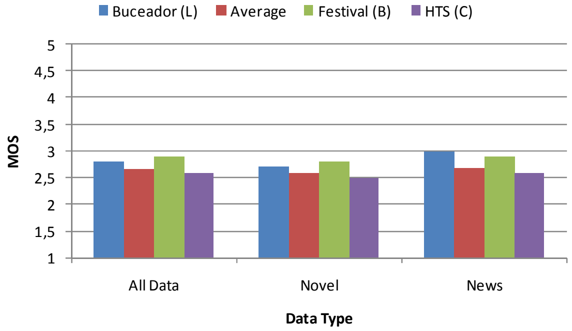

Figure 2: *Similarity to the Original voice*

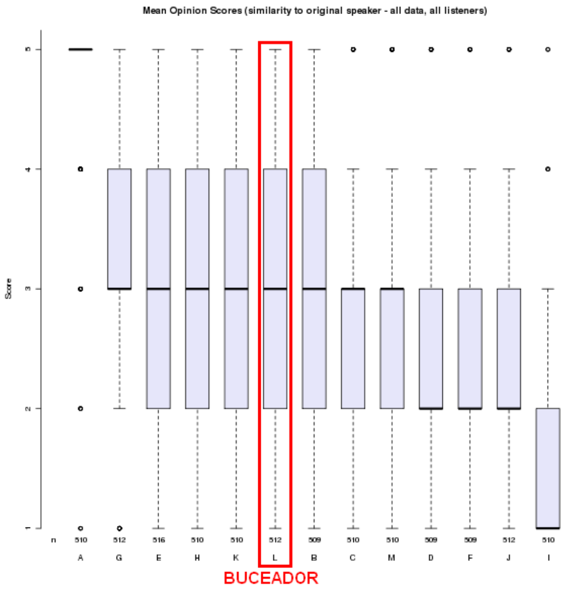

Figure 3: *Similarity boxplot for all data – all listeners*

Buceador TTS gets a similarity MOS of 2.8, whose absolute value is somehow low, considering that our system concatenates natural segments with almost no modification. We suppose that 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, our system scores above the average being significantly more similar to the original voice than 5 systems (C,D,F,J,I) and significantly less similar than 2 systems (G,E). Besides, there is no significant difference between our hybrid system and the reference

Festival system, whereas in Blizzard Challenge 2009, the reference system was significantly better in this task.

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*). Results are displayed in Figure 4 and Figure 5.

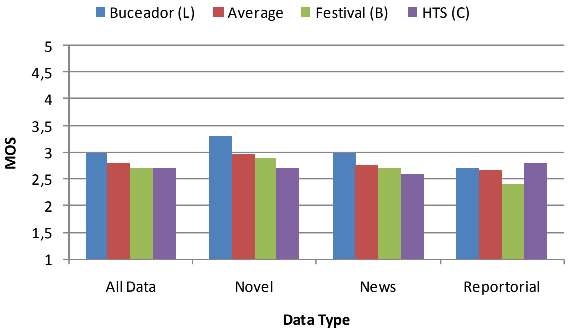

Figure 4: *Naturalness*

The Buceador system obtains a naturalness MOS of 3.0. It is significantly more natural than 7 systems (B,C,M,D,F,J,I) and significantly less natural than 3 (G,E,H). We believe 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 into the unit selection algorithm has alleviated this problem. Besides, the combination of two prediction methods produces a more robust prosody. While in Blizzard 2009 the reference system B was significantly better than ours, results have reversed this year, showing the performance improvement due to the hybrid approach.

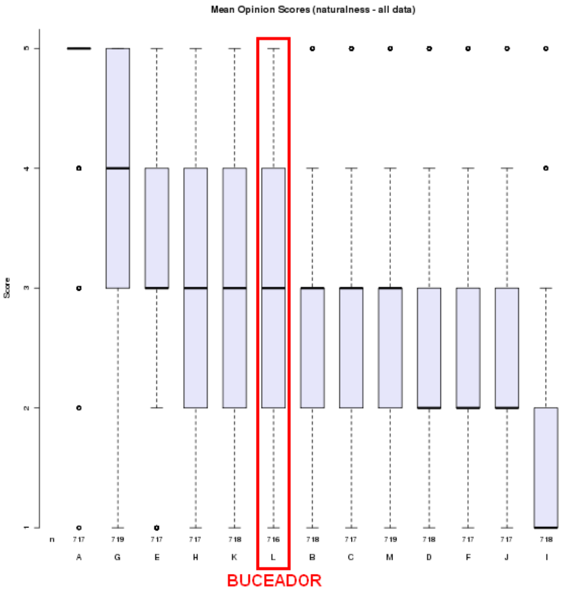

Figure 5: *Naturalness boxplot for all data – all listeners*

Looking at the genre of the texts synthesized, the worst results are obtained for the Reportorial data. That type of data consisted of long sentences in which the probability of having a bad join increases dramatically. In fact, the absolute value of the correlation between the average number of syllables per sentence in each domain (9.5 for Novel, 13.7 for News and 44.9 for Reportorial) and the MOS of Buceador, is pretty high: ρ=-0.92. We are not stating a causal relation between sentence length and MOS, because other issues must be taken into account as well (e.g. domain). But it is reasonable to expect that longer sentences will demand more complex phonetic or prosodic contexts that may have little or no representation in the corpus, thus reducing the quality of the synthesized speech.

4.3. Intelligibility Test

Organizers computed Word Error Rates (WER) for *SUS* and *addresses* as a measure of intelligibility. Figure 6 and 7 display these results.

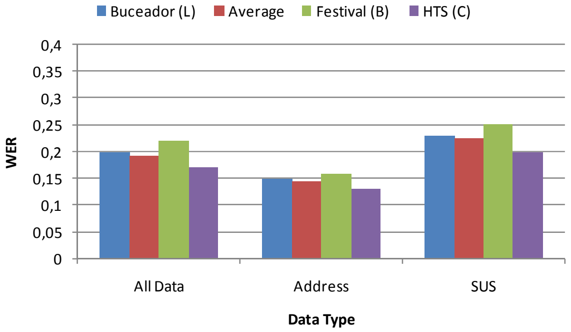

Figure 6: *Intelligibility*

Buceador TTS has a WER of 15% for *addresses*, 22% for *SUS* and 20% for *all data*. Two systems (C,F) are significantly more intelligible than ours, and there are no statistical differences with the rest. The speech submitted to the SUS intelligibility task was synthesized with the hybrid TTS, whereas the address task was synthesized with the HMM-based system. Our statistical system has yielded good intelligibility results, although there seems to be no significant differences among all the systems (including natural speech) in this particular task, perhaps due to its relative ease. The hybrid approach has taken advantage of the robustness of the statistical modeling, reducing the problems associated with labeling errors or poor pronunciations.

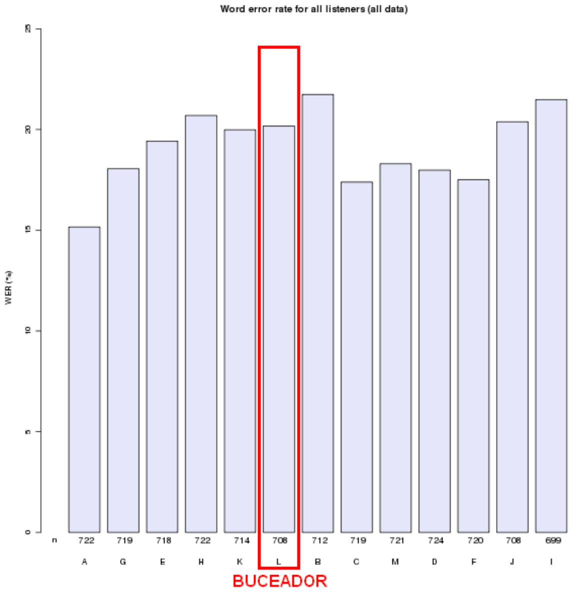

Figure 7: *Word error rate for all data – all listeners*

5. Conclusions

Two synthetic voices have been built for Blizzard Challenge 2011. On 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). Buceador system has been a joint effort between two research groups: UPC (linguistic processing) and Aholab (prosody and acoustic modules).

Although the obtained results have been quite good, several decisions or intrinsic circumstances may have prevented our system from achieving a better performance. Due to the lack of time and uncertainty about how long would it take to train the HMM-based system with such a big database, we did not modify our system in order to use the prosodic labels provided by Lessac (e.g. prominence and prosodic breaks). Their inclusion would probably improve the naturalness of the synthetic voice. In addition to this, the absence of native speakers in the development team hinders the necessary fine tuning of the voice. In any case, the fact is that the hybrid approach has reached promising results, significantly above average as far as similarity and naturalness are concerned, and with no significant differences with most of the systems in the intelligibility task.

There is still a significant gap between synthetic systems and natural speech in all the sections but the intelligibility task, in which the HMM-based system C gets comparable WER to the natural speech.

6. Acknowledgements

The authors would like to thank the organizers of Blizzard Challenge 2011 and the developers of all the tools employed during the voice building process.

This work has been partially supported by the Spanish Ministry of Science and Innovation (Buceador Project, TEC2009-14094-C04-00).

7. References

[1] A. Bonafonte, P. Agüero, J. Adell, J. Perez, and A. Moreno, "Ogmios: The UPC text-to-speech synthesis system for spoken translation," in *TC-STAR Workshop on Speech-to-Speech Translation*, 2006, pp. 199-204.

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

[3] A. Bonafonte, A. Moreno, J. Adell, P.D. Agüero, E. Banos, D. Erro, I. Esquerra, J. Pérez, T. Polyakova , "The UPC TTS system description for the 2008 blizzard challenge," *Proc of the Blizzard Challenge, 2008*, pp. 1-6, 2008.

[4] I. Sainz, D. Erro, E. Navas, I. Hernáez, I. Saratxaga, I. Luengo, I. Odriozola., "The AHOLAB Blizzard Challenge 2009 Entry," in *Blizzard Challenge 2009 workshop*, 2009, no. 2.

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

[6] I. Sainz, D. Erro, and E. Navas, "A Hybrid TTS Approach for Prosody and Acoustic Modules," in Proc. of *Interspeech 2011*, 2011.

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

[8] H. Fersøe, E. Hartikainen, H. Van Den Heuvel, G. Maltese, A. Moreno, S. Shammass, U. Ziegenhain, "Creation and Validation of Large Lexica for Speech-to-Speech Translation Purposes", in Proc. of *LREC*, May 2004.

[9] A. Bonafonte, H. Höge, I. Kiss, A. Moreno, U. Ziegenhain, H. van den Heuvel, H.-U. Hain, Xia S. Wang, and M.-N. Garcia. "TC-STAR: Specifications of language resources and evaluation for speech synthesis". In Proc. of *LREC*, pages 311–314, May 2006.

[10] 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," in *Proc. of Fala2010*, 2010, pp. 67-70.

[11] I. Sainz, D. Erro, E. Navas, I. Hernáez, J. Sánchez, I. Saratxaga, I. Odriozola, I. Luengo et al., "Aholab Speech Synthesizers for Albayzin2010," In *Proc. of Fala2010*, 2010, pp. 343-348.

[12] D. Erro, I. Sainz, E. Navas, and I. Hernáez, "HNM-Based MFCC+f0 Extractor Applied to Statistical Speech Synthesis," in Proc. of *ICASSP 2011*, 2011, pp. 4728-4731.

[13] A. Lazaridis, I. Mporas, T. Ganchev, G. Kokkinakis, and N. Fakotakis, "Improving phone duration modelling using support vector regression fusion," *Speech Communication*, vol. 53, no. 1, pp. 85-97, Jan. 2011.

[14] T. Hirai, J. Yamagishi, and S. Tenpaku, "Utilization of an HMM-based feature generation module in 5 ms segment concatenative speech synthesis," in *Proc. of the IEEE Speech Synthesis Workshop*, 2007, pp. 81-84.

[15] R. Nitisaroj, R. Wilhelms-tricarico, B. Mottershead, J. Reichenbach, and G. Marple, "The Lessac Technologies System for Blizzard Challenge 2010," in *System*, 2010.

[16] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. C. Woodland, "The HTK Book, version 3.4," 2006.

[17] I. Luengo, I. Saratxaga, E. Navas, I. Hernáez, J. Sanchez, and I. Sainz, "Evaluation of Pitch Detection Algorithms Under Real Conditions," in Proc. of *ICASSP'07*, 2007, p. IV-1057-IV-1060.

[18] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495. Elsevier, pp. 495-518, 1995.

[19] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]. Version 5.1.38.," 2010. [Online]. Available: http://www.praat.org/.