The AHOLAB Blizzard Challenge 2009 Entry

Iñaki Sainz, Daniel Erro, Eva Navas, Inma Hernáez, Ibon Saratxaga, Iker Luengo, Igor Odriozola

Aholab – Dept. of Electronics and Telecommunications.
University of the Basque Country. Urkijo zum. z/g 48013 Bilbo
inaki, daniel, eva, inma, ibon, ikerl, igor@aholab.ehu.es

Abstract

This paper describes the process of building unit selection voices for our participation in the Blizzard Challenge 2009. Out of the three voices required (EH1: 15 hours UK English, EH2: 1 hour UK English subset and MH: 6000-utterance Mandarin Chinese) only the English ones were built. As far as the Hub Tasks is concerned, only the ES1 task was completed using voice conversion techniques. The Evaluation show that some improvements have been made over last year results, especially for EH2 voice.
Index Terms: speech synthesis, unit selection, speech evaluation.

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 8 weeks for voice building, participants are asked to synthesize several hundreds of test texts that will be evaluated with respect to naturalness, similarity to the original speaker and intelligibility.

Aholab Signal Processing Laboratory has been developing since 1995 a complete TTS system for Basque and Spanish languages [1] using different acoustic engines: PSOLA, MBROLA [2], HNM and Corpus-based Unit Selection. This has been our second participation in the Blizzard challenge and various improvements have been made in our system since last year.

This paper is organized as follows. First, we describe the system with some detail, focusing on prosodic and acoustic modules. 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. System Overview

AhoTTS is the synthesis platform for commercial and research purposes of Aholab Laboratory. The system has a modular architecture, and written in C/C++ it is fully functional in Unix and Windows operating systems. In figure 1 we can see the system presented to the Blizzard Challenge 2009.

2.1. Text Normalization

Our efforts have focused mainly in the development of a complete TTS for both Basque and Spanish languages. In order to participate in the Blizzard Challenge 2009 we have used *Festival* [3] as the text processing module for English.

To make the communication between Festival and AhoTTS possible, we have chosen the XML inter-module interface for synthesis systems specified by the ECESS [4]. Being the sentence hierarchy of ECESS very similar to the "*Utterance*" of Festival, the format conversion has been quite straightforward once POS and internal phone-set were properly mapped to ECESS format and Sampa respectively.

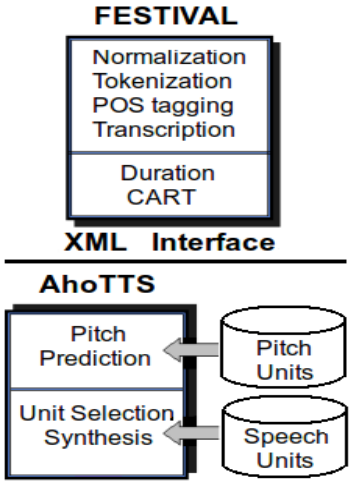

Figure1: *System Overview.*

2.2. Prosody Prediction

In Blizzard 2008 edition we were not able to adapt our corpus-based pitch contour prediction to English due to time constraints. But for this year we have developed speaker dependent models for each of the English Voices.

2.2.1. Duration Model

CART zscore duration models were trained using the *wagon* tool, the provided speech data and the features suggested in [5].

2.2.2. Pitch Model

Our unit selection intonation modelling for Basque and Spanish has been adapted to English as well. We use the voiced phoneme as the basic unit size in a similar approach to [6]. Such a small unit provides greater flexibility, although the concatenation of not consecutive units within a syllable 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 target and concatenation sub costs as shown below:

$$C(t_1 \ldots t_n, u_1 \ldots u_n) = \alpha \sum_{i=1}^{n} C^T(t_i, u_i) + (1-\alpha) \sum_{i=1}^{n-1} C^C(u_i, u_{i+1}) \quad (1)$$

$$C^T(t_i, u_i) = \sum_{j=0}^{P} w_j^T C_j^T(t_i, u_i) \quad (2)$$

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

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-

10.21437/Blizzard.2009-2

th weight of the P target sub costs and the Q join sub costs. The main features employed in the target function are listed below:

- *Type of proposition:* Declarative, interrogative, exclamatory, unfinished, etc.
- *Segmental characteristics* (phoneme type, articulation point, voicing, etc.) of the neighbouring phonemes.
- *Position* (single, start, middle, end) in the syllable, in the word, in the phonic group, etc.
- *Relative position* to the accented syllable and to the nearer pause.
- *Duration* of the neighbouring phonemes compared to the predictions of the CART model.
- *Simplified POS* of the word in which the unit is included.

Target weights are adjusted using a similar approach to the one proposed in [7] 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 with the summation of the target sub costs defined above, solving the weights as a multiple linear regression problem.

The following join sub costs are employed when two units are not consecutive in the corpus:

- *Pitch difference at the join* when both are voiced units without intermediate unvoiced phonemes.
- *Pitch difference among natural neighbours of the units to be concatenated*: next natural contour of the current unit compared to the previous natural contour of the next candidate unit, and vice versa.

Join weights are manually assigned and some penalization are added in order to hinder the concatenation of non consecutive voiced units inside a syllable, and to a lesser level, inside a word. Finally the intonation curve is slightly smoothed in the joins, interpolated in the unvoiced parts and scaled to the desired phone duration.

2.3. Acoustic Engine

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

We use half-phones as the elementary unit because of the flexibility they provide to form longer units. From the objective phone sequence, context-dependent half-phones are generated. If there are enough candidates in the database for a specific context (more than a threshold), we generate diphone units because they preserve the coarticulation effect and the concatenation in the stable part of a phone is usually less problematic. On the contrary, if sufficient candidates cannot be found, all the half-phone contexts for that phoneme are added to the list.

2.3.1. Unit Selection Algorithm

Target cost function (2) is divided in the following sub costs which are applied at the demiphone level:

- *Phoneme context:* A discrete value cost within a 7 phoneme window and with the following features:.Phoneme identity, Vocal/Consonant, vowel height, vowel frontness, vowel roundness, vowel length, Voiced/Unvoiced, consonant articulation mode and consonant articulation point.
- *Pitch:* Euclidean distance of pitch contours sampled each 5ms with a previous normalization of the duration.
- *Pitch Slope*: Pitch difference in a small window at the extremes of the units which are adjacent to a voiced phoneme.
- *Duration:* difference in unit length. For voiced units the number of pitch marks are taken into account (small duration modifications can be applied with little quality lost).
- *Accent:* Relative distance to the nearest accented syllable because units before and after an accent may have different characteristics.
- *Type of proposition:* Declarative, interrogative, exclamatory, unfinished, etc.
- *Relative position in the phonic group* (interval within pause breaks).
- *Voiceness*: It penalizes voiced phones marked as unvoiced during the pitch detection process.
- *Position in the syllable and in the word:* Single, start, middle, end.
- *Score*: It tries to measure the pronunciation quality of the unit. It is precomputed as the normalized spectral distance to the centre of each halfphone cluster.

The concatenation cost function (3) is composed of seven sub costs, all but the *inter-syllable pitch range* being calculated only for non-consecutive units.

- *Pitch:* Pitch difference at the concatenation point.
- *Pitch Slope:* Pitch slope difference at the concatenation point within voiced units.
- *Inter-half-phone pitch range:* If the difference between the maximum and minimum pitch values of two adjacent voiced units exceeds a threshold, the join is penalized. The threshold is calculated from the natural values of the database for each of the possible phoneme class combinations.
- *Inter-syllable pitch range:* To control excessive pitch jumps among consecutive syllables. The threshold is database dependent and divided into 3 groups depending on the stress during the syllable transition: (i) both syllables are stressed, (ii) only one and (iii) none of them.
- *Duration:* The difference between the objective duration of the whole phoneme and the sum of intra-phoneme half-phones.
- *Power:* Energy difference between last and first frame, and the overall energy too for intra-phoneme voiced half-phones.
- *Spectrum:* Euclidean distance between vectors of 13 MFCC coefficients with delta and acceleration values. The result is normalized with the precomputed mean distance of the transitions of the natural units from the corpus. Those distances are stored separately for each phoneme if they are intra-phoneme transitions, and clustered by phoneme type for inter-phoneme ones.
- *Voiceness:* Penalizes the join between non-consecutive units detected as unvoiced, because its pitch marks can be less reliable.
- *Penalizations*: Concatenations in the transition between phonemes are hindered in favour of the concatenations in the stationary part (middle of a phoneme). Several penalizations are deployed depending on the voiceness and articulation point of the phonemes. Besides, consecutive joints are favoured in intra-syllable transitions while inter-word and inter-syllable transitions are less penalized.

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 sub costs as the predictors. Different weights are estimated for left and right halfphones and for each phoneme type.

Concatenation weights are adjusted manually given more importance to the pitch and spectral distances than to the energy. In the same way, α weight from equation (1) is smaller than 0.5 in order to boost the concatenation cost over the target one.

2.3.2. Waveform Generation

The candidate units selected 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 modifications are executed related with changing the duration of the voiced signal by means of pitch synchronous overlap and add techniques. The energy is also 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. Building the Blizzard Voices

The English data set provided is the same as last year. It was recorded at CSTR and comprises 15 hours of speech recorded by a male speaker with southern British accent. The dataset is composed of data from different genres: Dialogue rich children stories (1390 utterances), isolated words (2880 utterances), CMU Arctic (1132 utterances), carrier sentences for emphasized words (1681 utterances) and newspaper texts (2449 utterances). The recordings are supplied as mono waveform files, with 16kHz sample rate and 16 bit precision.

The whole process explained in the following subsections has been applied to the full database (EH1) and to the 1 hour *arctic* subset of it (EH2). ES1 task, in which only 100 sentences of the *arctic* subset could be used, is briefly explained in a specific subsection.

3.1. Segmentation

Due to limited time and the huge amount of data provided it is impossible to check whether the text transcriptions match with what actually the speaker is saying or not. So, only some upper-case words were revised to discover if the speaker has spelt them or pronounce them as expected. Moreover, some out of vocabulary words were added to the Unilex lexicon.

As we have no acoustic models for English, a forced alignment process has been implemented in order to obtain the segmentation labels. HTK [8] toolkit has been employed within the script provided in the *multisyn* building package [9] and with the transcription labels extracted from Festival utterances. During the alignment, vowel reduction is set as an alternative phone substitution and in fact, many schwas are inserted in the intermediate segmentation. As a post-processing, the "reducable" feature extracted from the festival utterances has been used in order to maintain only the more probable schwas. As an optional short pause "sp" model between words is used during the alignment, pauses shorter than a specified threshold are removed and waveform normalization is performed in all the signal. Finally, after all the post-processing is done, a last forced alignment pass is realized.

Once labelling is completed, we convert the unilex internal phone-set of Festival to Sampa (adding some special phonemes from unilex set). After a quick analysis we can conclude that the quality of the segmentation is worse than expected, so an intense pruning of the data has been made. Data to be pruned is selected by means of the alignment scores from *HVite*, spectral distance to the centre of the cluster for each phoneme and the detection of extreme duration outliers. No manual correction has been done to any of the labels.

3.2. EH1 and EH2 Voice Building

The EH1 database contains several genre and styles that offer a great variability to the voice but cause troubles when units from completely different sessions are mixed together. As we have not obtained proper models for some of the genres and did not have automatic genre detection (from a plain text input), up to 41.7% of the database was dropped for EH1 voice building: conversation (13%), unilex (28%), address (0.4%) and spelling (0.3%).

After an initial genre pruning for EH1, the following voice building steps were applied separately but the same way to EH1 and EH2 English voices. Power normalization is performed measuring the mean power in the middle of the vowels, and normalizing each inter-pause interval which that value. Then, pitch curve is detected with our own PDA (Pitch Detection Algorithm) [10] based on cepstrum and dynamic programming. Pitch marks are then generated with another tool designed in our laboratory: a quite simple peak marking that relies on the accuracy of our PDA, and interpolates the marks in the unvoiced parts. Edinburgh speech tool *sig2fv* is used to generate 13 MFCC parameters calculated with a fixed 5ms frame. For each unit the following information is stored:

- *Power:* Log power values in the extremes and the middle of the unit.
- *Pitch:* 3 point linear curve with the first, last and the most significant point.
- *Spectrum:* MFCC, delta an acceleration coefficients for the first and last frame.

Apart from that, all the necessary linguistic information is extracted from the Utterance structure of Festival and merged with the rest of the data in a single binary file. Finally, prosody models and target weights are trained from the generated features.

3.3. ES1 Voice Building

An alternative system has been used in the ES1 task. In order to add voice conversion capabilities to the AhoTTS system, an attempt has been made to combine its waveform generation block with an acoustic module based on the algorithms and methods presented in [11], according to the cascade architecture shown in figure 2.

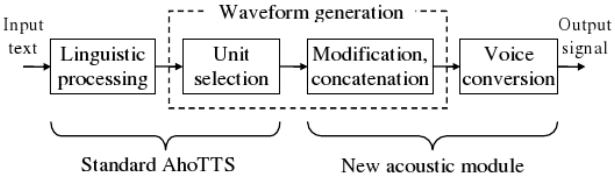

Figure 2: *Architecture of the synthesis + conversion system under evaluation.*

AhoTTS performs the text processing, prosody generation and unit selection tasks, whereas the appended acoustic module modifies the pitch and duration of the selected units, concatenates them, and applies voice conversion techniques to the resulting signal, using a Harmonic plus Stochastic speech Model (HSM). Voice conversion is performed at two different acoustic levels:

- Pitch adaptation, which consists in replacing the mean and variance of the source log-f0 distribution by the values measured from the target voice.

- Spectral conversion using the weighted frequency warping (WFW) technique. WFW applies a time-varying frequency warping function combined with a correction filter given by typical GMM-based linear transformations (see [11] for details).

AhoTTS communicates with the new module through a very simple interface that adapts the output files of the synthesizer to the input requirements of the HSM-based acoustic module. They are kept as independent tools and no specific work has been carried out to integrate both systems in an optimal way.

3.3.1. Training and Corpora

The default voice of the TTS system was built automatically from a UK database called Laura, which was recorded within the scope of the TC-STAR project (IST- FP6-506738, funded by the European Commission). The database belongs to Siemens AG, Munich, and it consists of approximately 10 hours of speech from one female speaker. Among the few UK-English databases available for our group, Laura was the only one suitable for synthesis using AhoTTS. Therefore, it has to be emphasized that no special care was taken about the suitability of Laura voice for voice conversion.

Due to the time constraints (the whole system was prepared in one day), as training specific duration models was time-consuming, generic duration models were used instead. Some differences were found between the phoneset of Roger and that of Laura. This problem was overcome using a manual mapping of phones, although some phones in Laura's phoneset had to be paired with more than one phone in Roger's phoneset, resulting into a small loss of phonetic information.

In order to train the spectral voice conversion function on a parallel corpus, Roger's training sentences were generated using the synthesizer (with Laura's voice). The resulting pseudo-parallel sentences were aligned in time via piecewise linear time-warping functions, which were defined from reference instants placed at the phoneme boundaries. The source-target frame pairs were used to train the GMM from whose parameters the spectral transformation function was defined. The number of Gaussian mixtures of the statistical model was configured manually: 8th order for the 100-sentence training set, 4th order for the 50-sentence training set, and 1st order for the 10-sentence training set.

4. Evaluation

For each voice, participants were asked to synthesize hundreds of sentences from 4 genres: conversational speech (conv), semantically unpredictable sentences (sus), texts from stories (novel) and news (news).

Listeners were divided in 3 groups during the web-based evaluation: Paid participants (EU, all native speakers of English), Volunteers (ER) and Speech Experts (ES). Each group performed six evaluation tasks: (i) Mean Opinion Score (MOS) to measure the similarity with the original voice, (ii) Similarity test between two voice samples, Two MOS naturalness tests with (iii) conversational domain sentences and (iv) news, (v) an intelligibility test in which listeners were asked to transcribe the SUS they heard and (vi) MOS appropriateness in conversational domain.

4.1. Results for voices EH1 - EH2

More than four hundred subjects took the evaluation test. The final results are commented in the following section comparing our performance with the other participants. It must be stressed that natural voice (system A) was presented just as another system in order to establish the ceiling score. Besides, some benchmark voices participated in the challenge:

- *Standard Festival unit selection:* system B, voice built using the same method as used in the CSTR entry to Blizzard 2007.
- *Standard speaker-dependent HMM-based voice:* system C, built using a similar method to the HTS entry to Blizzard 2005.

These reference voices are of great value in case we want to study the evaluation results from different years. Since Blizzard Challenge 2009 uses the same English database that the 2008 evaluation, it should be easier to notice improvements of the participants.

4.1.1. Naturalness Test

The naturalness MOS scores from 1 (sounds completely unnatural) to 5 (sounds completely natural). A comparative graphic among the benchmark systems (Festival and HTS), the average of all participants (Avg, system A excluded) and the results of our system (for all listeners and for ER, EU and ES groups) is shown in Figure 3. Besides, it displays both current and last year results. The same information for voice EH2 can be seen in figure 4.

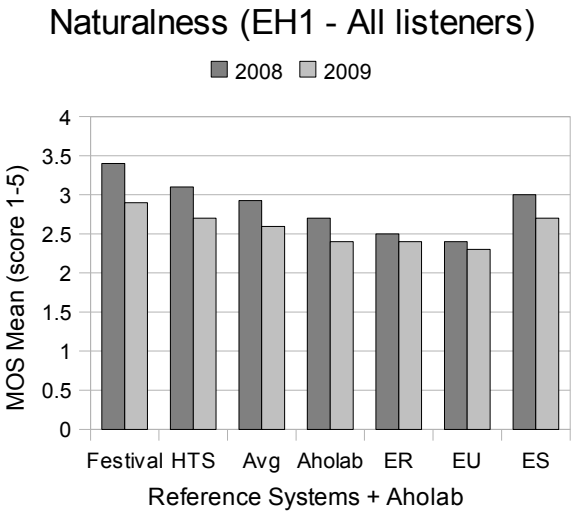

Figure 3: *Naturalness for Voice EH1.*

Benchmark systems Festival and HTS have had a mean drop of 14.71% (from 3.4 to 2.9) and 12.90% (from 3.1 to 2.7) respectively. The average of all participants (20 last year, 17 in 2009) has fallen 11.31% (from 2.93 to 2.59), while our system gets a slightly smaller drop of 11.11% (from 2.7 to 2.4). Therefore, even if our mean MOS is clearly lower than the previous year one, comparison with the reference systems indicates that there has not been a declination in performance.

As far as the *Pairwise Wilcoxon signed rank* results is concerned, 7 systems score significantly higher than us, 7 system significantly lower and 2 (D and O) equally. Last year only 2 systems scored significantly lower than Aholab and 8 higher. That seems an improvement over last year results, although it must be stated that the number of participants has not remained constant.

Not surprisingly, we score considerably better for the news domain texts (2.7 mean MOS instead of 2.4). This results were quite expected bearing in mind that 55.82% of the data used for voice EH1 after pruning, consisted of texts from The Herald Newspaper.

It must be emphasized that the average score for EH2 beats the one of EH1, 2.62 and 2.59 respectively. Our system also obtains better results for EH1 (2.6) than for EH2 (2.4)

even if former database is almost 8 times bigger (after pruning) and EH2 is a subset of EH1. The greater variability of the EH2 can be a possible explanation for that degradation in performance. EH2 contains multiple sessions, genres and styles that some systems seem to cope with worse than others.

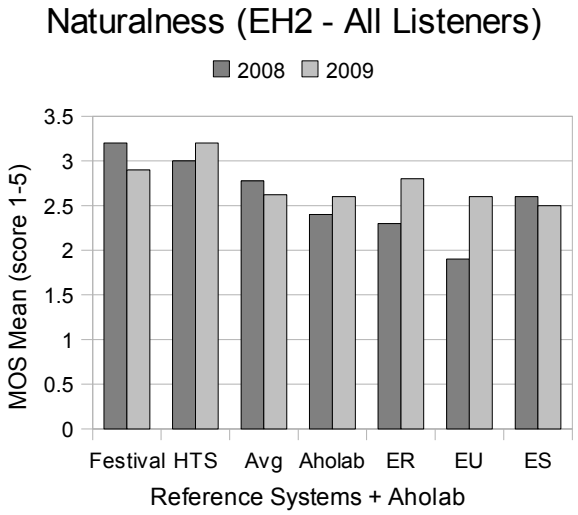

Figure 4: *Naturalness for Voice EH2.*

4.1.2. Similarity Test

MOS (1: sounds like a totally different person; 5: sounds exactly as the same person) comparative among the benchmark systems and our system is shown in figure 5 for EH1 voice and in figure 6 for EH2.

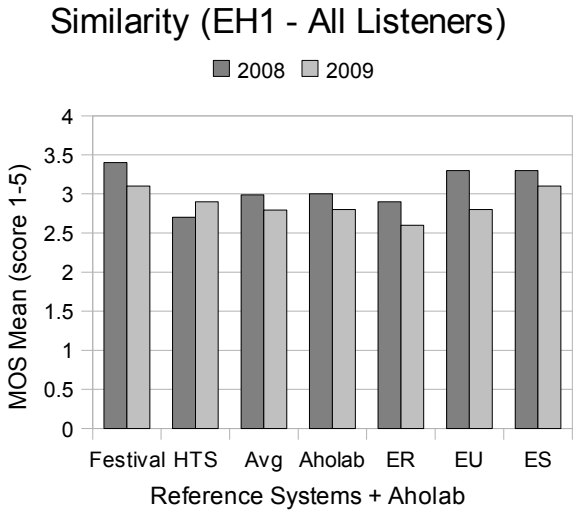

Figure 5: *Similarity for Voice EH1.*

The similarity of EH1 to the original voice has fallen from 3.0 in 2008 to 2.8 in 2008 (-6.66%) for our system, -8.82% for Festival, -6.55% for the average system and a 7.40% rise for HTS. While the Festival reference system scored significantly better than Aholab last year, both systems do not seem to have significant differences now.

As far as EH2 voice is concerned, we get exactly the same mean MOS than in 2008 (2.7), while the average system drops -8.43%, -3.12% Festival and HTS rises 20.00% its performance. The wilcoxon test brings similar results as the EH1 voice: Festival benchmark system is not significantly better anymore, and last year it was. Being our system a concatenative one, we could expect better similarity results than statistical methods as the reference HTS voice, since we use natural speech units. An explanation for this performance could be acoustic artifacts that sometimes appear in concatenative systems without spectral smoothing, and make focusing only on similarity more difficult for listeners.

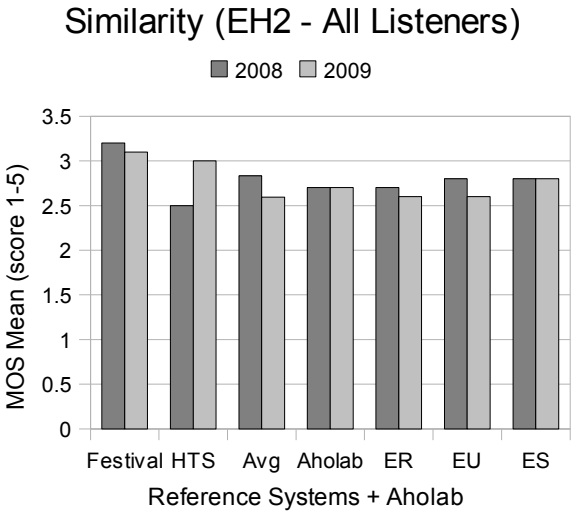

Figure 6: *Similarity for Voice EH2.*

4.1.3. Word Error Rate Test

Figure 7 and 8 show the WER (Word Error Rate) for benchmark systems and ours, for EH1 and EH2 respectively. Only the responses of English native listeners are taken into account because we consider that non native listeners add mainly noise to the evaluation of this section.

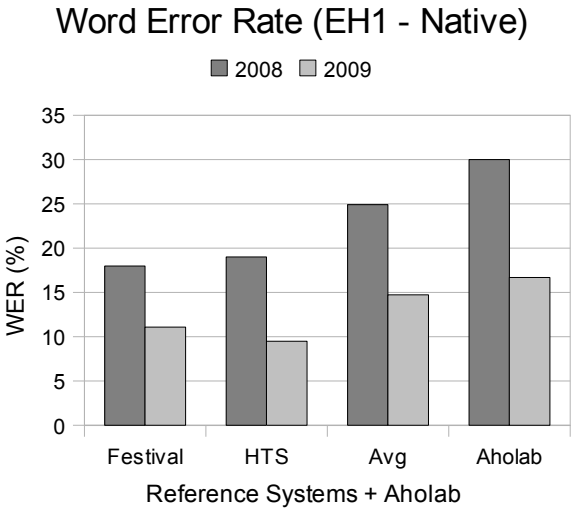

Figure 7: *WER for EH1.*

This year, the organization used a new SUS generator provided by Tim Bunnell, which uses less complex words than in previous challenges. This must be the main explanation to the incredible improvement in the intelligibility of the systems compared with last year. Aholab gets 16.7% WER for EH1, which represents a reduction of 44.33% over last year, a greater improvement than average system (40.88% reduction) and Festival (38.33%). We believe this relative improvement is due to better outlier detection and to the score target function in the unit selection algorithm. Surprisingly we obtain better results for EH2 voice (15.82% WER), and the same can be said for the average system: 14.33%.

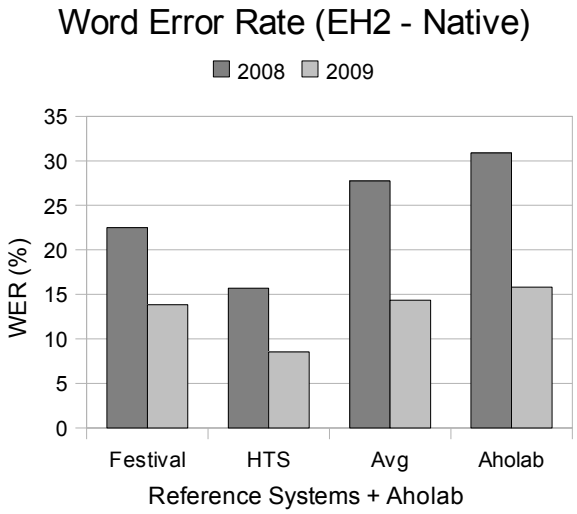

Figure 8: *WER for Voice EH2.*

4.2. Results for Hub Task ES1

The system was built and trained in only one day, so any observation about the results should take that into account. Although both the synthesizer and the HSM-based acoustic module had given good results in other evaluations, the performance of the combined system is far from being acceptable. Obviously, the main reason is the very little time spent building the system. In fact, the system consisted in connecting two independent sub-systems without any special care about their interaction. This unsupervised interaction (which takes place not only during synthesis but also during training) seems to cause a significant quality loss in the synthetic signals with respect to the original signals. Therefore, the naturalness scores are low. Moreover, the voice of the synthesizer was very different from Roger's voice, mainly in terms of gender and recording conditions. The voice conversion technique applied to transform the former into the latter, WFW, did not succeed at compensating such differences. In principle, we expected to obtain more interesting results when testing the system under such adverse conditions (very little time for training, very different source voice). However, we conclude that there are some aspects whose relevance should not be underestimated, especially the careful selection of an adequate voice for the system.

5. Conclusions

This has been our second participation in the Blizzard Challenge and the evaluation results show an improvement over last year (especially for the smaller EH2 voice). We still believe that the segmentation is one key issue we should focus on to obtain a better performance. Instead of forced alignment with a flat start, an initial segmentation provided by a DTW between natural recordings and a diphone synthesizer of Festival voices, could led us toward less segmentation errors.

Not having any English native speaker in our laboratory is another disadvantage. It hinders the tuning of the voice, so that we have used almost the same manually set concatenation weights that we had for Basque Language.

Probably, we will have to make a decision about our future approach: whether to use statistical synthesis, HSM waveform generation that allows spectral and prosodic modifications, or some kind of hybrid.

We have found this international evaluation to provide a good opportunity and stimulation to improve the quality of our system. Therefore, we are willing to participate in future campaigns as well.

6. Acknowledgements

The authors would like to thank to all people who supported and organized the Blizzard Challenge 2009, and also to the developers or owners of the various tools and databases employed during the voice building.

7. References

[1] Hernáez, I., Navas, E., Murugarren, J.L., Etxebarria, B.: "Description of the AhoTTS system for the Basque language", In SSW4, paper 202, 2001.

[2] Etxebarria, B., Hernáez, I., Madariaga, I., Navas, E., Rodríguez, J. C., Gándara, R. "Improving quality in a speech synthesizer based on the MBROLA algorithm." Proc. Sixth European Conference on Speech Communication and Technology, pp. 2299-2302, Budapest, 1999.

[3] Taylor, P., Black, A. and Caley, R. "The architecture of the Festival Speech Synthesis System", 3rd ESCA Workshop on Speech Synthesis, pp. 147-151, Jenolan Caves, Australia, 1998

[4] Javier Pérez, Antonio Bonafonte, Horst-Udo Hain, Eric Keller, Stefan Breuer, and Jilei Tian "ECESS inter-module interface specification for speech synthesis" Proceedings of LREC Conference, 2006

[5] Black, A.W. and Lenzo, K.A., "Building Synthetic Voices, Language Technologies" Institute, Carnegie Mellon University and Cepstral. LLC, 2003.

[6] Raux, A., Black, A., "A unit selection approach to f0 modeling and its application to emphasis", Proc. of ASRU 2003, St Thomas, US Virgin Is, 2003.

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

[8] 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.

[9] Robert A. J. Clark, Korin Richmond, Simon King, "Multisyn Voices from ARCTIC Data for the Blizzard Challenge", in INTERSPEECH-2005, 101-104, Lisboa, Portugal, 2005.

[10] Luengo, I., Saratxaga, I., Navas, E., Hernáez, I., Sanchez, J., Sainz, I. (2007): "Evaluation Of Pitch Detection Algorithms Under Real Conditions." Proc. of 32nd IEEE ICASSP, pp. 1057-1060, Honolulu, 2007.

[11] D. Erro, A. Moreno, A. Bonafonte, "Flexible Harmonic / Stochastic Speech Synthesis", 6th ISCA Workshop on Speech Synthesis, 2007.