The AHOLAB Blizzard Challenge 2008 Entry

Iñaki Sainz, Eva Navas, Inma Hernáez

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

Abstract

This paper describes the process of building unit selection voices for our participation in the Blizzard Challenge 2008. Out of the three voices required (15 hours UK English, 1 hour UK English subset and 6.5 hours Mandarin Chinese) we only built the English ones.

Index Terms: speech synthesis, unit selection

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 7 weeks for voice building, participants are asked to synthesize several hundreds of test texts that will be evaluated with respect to naturalness 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 first participation in an international evaluation campaign, and also our first attempt to build an English synthetic voice.

First, we describe with some detail our system focusing on the acoustic module. 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.

2. System Overview

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

2.1. Text Normalization

As mentioned previously, our efforts have focused mainly in the development of a complete TTS system for both Basque and Spanish languages. Therefore, if we intended to participate in the Blizzard Challenge we needed a text processing module for English. Due to the huge work and the necessary knowledge to deploy such a module from scratch, we decided to make use of an already existing one. Of course, *Festival* [3] was our first choice.

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

2.2. Prosody Prediction

Our first intention was to employ our corpus-based pitch contour prediction, but due to time constraints we were not able to adapt it to the English voice. So, we relied once again in Festival.

CART duration and intonation models were trained using the *wagon* tool and the provided speech data. Finally, a scheme module was written in order to obtain the ECESS XML input for the acoustic module.

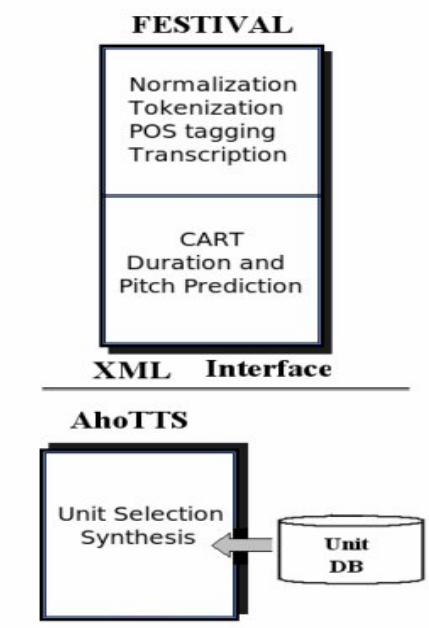

Figure 1: System Overview

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 concatenation 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 target phone sequence, context-dependent half-phones are generated. If there are enough candidates (more than a manually adjusted threshold) in the database for a specific context, 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, preselected half-phone contexts are added to the list. The selection of those contexts is done in advance during the voice building phase, by means of calculating the spectral cluster center of each context and annotating the nearest ones. If some context-dependent half-phone is missing in the corpus, alternatives are searched taking into account the context type (plosive, nasal, etc).

2.3.1. Unit Selection Algorithm

Our unit selection system implements a generic Viterbi search to find the sequence of candidate units from the database that minimizes a function cost composed by target and concatenation subcosts as shown below:

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

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

$$C^C(u_i,u_{i+1})=\sum_{j=0}^{Q}w_j^C C_j^C(u_i,u_{i+1}) \quad (3)$$

Where 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 concatenation subcosts.

Target cost function is divided in the following subcosts which are applied at the demiphone level:

- *Triphone:* A discrete value cost that favors the use of consecutive units from the corpus.
- *Context:* A discrete value to penalize half-phones with different left/right context.
- *Pitch:* Euclidean distance of pitch contours sampled each 5ms with a previous normalization of the duration.
- *Duration:* absolute difference in unit length. It takes into account if there are enough pitch marks in voiced units that will allow small duration modifications with little quality lost.
- *Accent:* Distance to the nearest accent measured in phonemes, because units before and after an accent have different characteristics.
- *Type of proposition:* Declarative, interrogative, exclamatory, unfinished, etc.
- Relative position in the sentence.
- *Adjacent phone type*: A two phoneme window is used.
- *Voiceness*: Penalizes voiced phones detected as unvoiced during the pitch detection algorithm, to avoid selecting possible poor pronunciations or units with wrong pitch marking.

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

- *Pitch:* Pitch difference at the concatenation point.
- *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.
- *Inter-syllable pitch range:* To control excessive pitch jumps. It is similar to the previous cost but calculated between consecutive syllables. The threshold is database dependent too.
- *Duration:* The difference between the objective duration 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 of two 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 the pitch marks may be less reliable.

Target weights are adjusted using a similar approach to the one proposed in [5]. We measure the spectral distance between units in the database and try to predict it with the summation of the target subcosts defined above solving the weights as a multiple linear regression problem.

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

2.3.2. Unit Concatenation

The candidate units selected are joined using glottal closure instant information to get smooth concatenations. It is well known that prosody modification hinders the overall natural quality of the voice. Therefore, only minor modifications are effectuated. These modifications are related with changing the duration of the voiced signal by means of pitch synchronous overlap and add techniques, and the modification of the energy contour of the units.

3. Building the Blizzard Voices

The English data set provided was recorded at CSTR and comprised 15 hours of speech recorded by a male speaker with southern British accent. The dataset was 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 were supplied as mono waveform files with 16kHz sample rate and 16 bit precision.

The whole process explained in the following subsections was applied to the full database and to the 1 hour subset of it.

3.1. Segmentation

Due to limited time and the huge amount of data provided, it was not possible to check manually whether the text transcriptions matched with what actually the speaker was saying or not. So, only some upper-case words were revised to discover if the speaker has spelt them or pronounce them as expected.

As we did have no acoustic models for English, a forced alignment process was implemented in order to obtain the segmentation labels. HTK [6] was employed within the script provided in the *multisyn* building package [7]. The phone

labels were extracted from Festival utterances. During the alignment, vowel reduction was set as an alternative phone substitution and in fact, many "schwa" were inserted in the final segmentation.

Once the labelling was completed, we convert the unilex internal phone-set of Festival to Sampa. The quality of the segmentation was worse than expected, so an intense pruning of 20% of the data was made. Data to be pruned was selected by means of the alignment scores from HVite and extreme duration outlier detection.

3.2. Voice Building

The following steps for voice building were fully automatic. Power normalization was performed measuring the mean power in the middle of the vowels and normalizing each inter-pause interval with that value. Then, pitch curve was detected with our own PDA algorithm [8] based on cepstrum and dynamic programming. Pitch marks were generated with the help of "epoch" tool from ESPS, limiting the pitch range with the results from our PDA algorithm, and interpolating the marks in the unvoiced parts. Edinburgh speech tools *sig2fv* was used to generate 13 MFCC parameters calculated with a pitch synchronous window. For each unit the following information was 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 significative point
- *Spectrum:* MFCC, delta and acceleration coefficients for the first and last frame.

Finally, all the linguistic information was extracted from the Utterance structure of Festival and merged with the rest of the data in a single binary file.

4. Evaluation

The evaluation for English consisted of two databases, 15 hour database (Voice A) and its ARCTIC subset 1 hour database (Voice B). For each voice, participants were asked to synthesize 620 sentences from 5 genres: conversational speech (conv), semantically unpredictable sentences (sus), sentences with emphasized words, text from stories (novel) and news (news).

Three categories of listeners were used in the web-based evaluation: (i) Paid students (British and Indian), (ii) Volunteers and (iii) Speech Experts. Each group performed five evaluation tasks: (i) Mean Opinion Score (MOS) to measure the similarity with the original voice, (ii) Similarity of the naturalness of two voice samples, two MOS tests with (iii) novel domain sentences and (iv) news, and (v) an intelligibility test in which listeners were asked to transcribe the SUS they heard.

4.1. Results

More than three hundred subjects took the evaluation test. The final results are commented in the following lines comparing our performance with that of the other participants. It must be stressed that natural voice (A system) was presented just as another system in order to establish the ceiling score and to detect "unwanted" listeners that answer randomly.

4.1.1. Similarity Test

The boxplots of similarity scores of all systems for voice A are shown in Figure 2. The 5-point scale scores from 1 (Sounds like a totally different person) to 5 (Sounds like exactly the same person). The central solid bar represents the median, the shaded box the quartiles, extended lines the 1.5 times quartile range, and the outliers are displayed as circles.

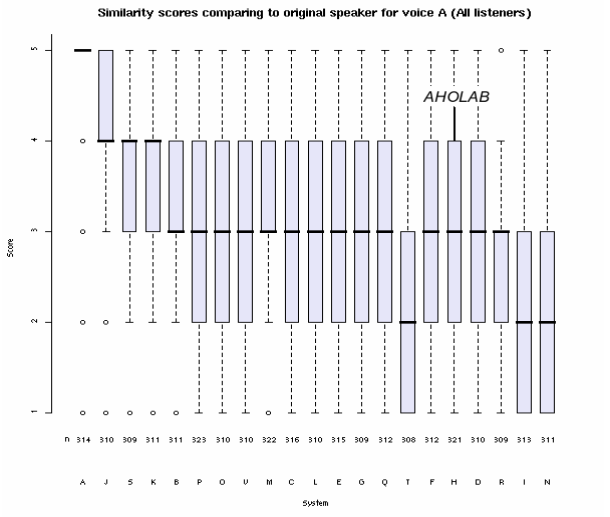

Figure 2. Boxplots of similarity test results for voice A

From the Figure 3 we can conclude that our system performs only slightly worse than the average of the rest of the systems (removing system A) in the similarity test.

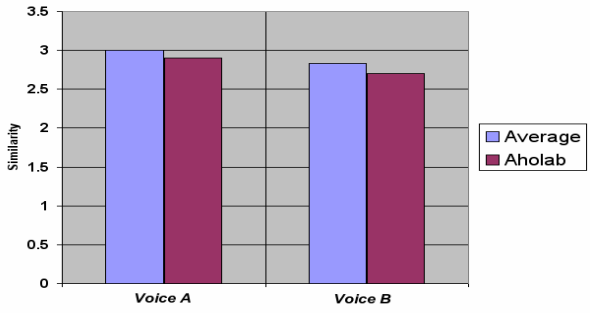

Figure 3. Similarity comparison among our system and the average system

4.1.2. Mean Opinion Score Test

MOS (1: sounds completely unnatural; 5: sounds completely natural) comparative between the average system and ours for all the listeners is shown in Figure 4.

The improvement in the results obtained by Voice A (2.7 MOS) was smaller than expected when compared with those of Voice B (2.6 MOS) that was 15 times smaller. This may be due to the bigger intersession variability and segmentation errors of the former database.

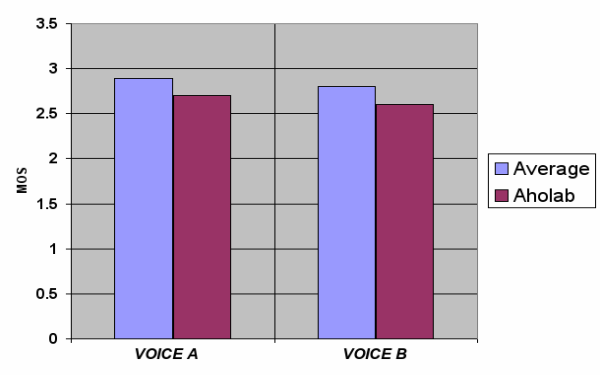

Figure 4. MOS comparison between our system and the average system

4.1.3. Word Error Rate Test

Figure 5 shows the word error rates for all the participants with Voice A. We score worse than expected in the intelligibility test with 44% word error rate for the Voice A, and 49% with Voice B. One of the possible explanations for such a high error can be the difficulties that non-native English speakers face when they have to properly tune their systems.

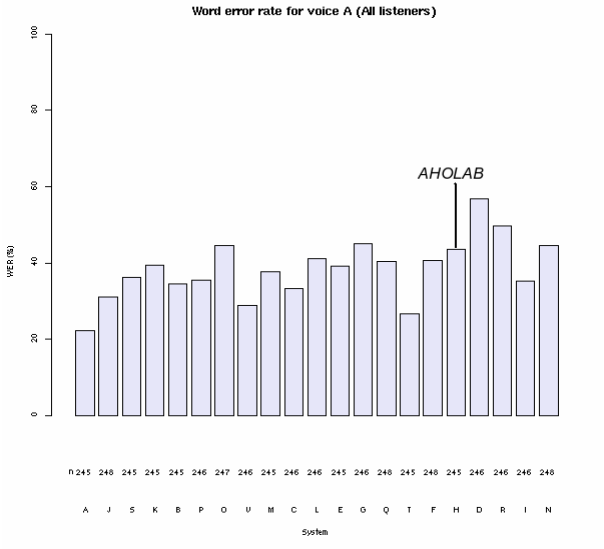

Figure 5. Word Error Rates for voice A

5. Conclusions

The Blizzard Challenge was the largest test that our system has participated in.

The listening tests carried out as part of the Blizzard challenge 2008 have shown that the resulting quality is near the average of the other systems. We are certain that there is much room for improvement in the quality of the synthesized voice, especially in the prosody prediction module. But given the fact that it was our first participation in the Blizzard challenge and out first English voice, we consider these results as promising.

We believe that one of the key factors that has worsened the performance of our system is the segmentation of the data, specially the schwa insertions. Due to a mismatch between the transcriptions gathered from the unilex lexicon through Festival and the phone labels of the database, our unit selection algorithm had difficulties finding proper signals. We should have trained "post-lexical vowel reduction rules" for Festival but it was not possible to do so in time. Besides, a corpus-based prosody prediction will have probably increased both the naturalness and the pleasantness of the synthesized voice.

In any case, 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 the Blizzard Challenge 2008, and also to the developers of the various tools or systems employed during the voice building process.

7. References

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

Etxebarria, B., Hernáez, I., Madariaga, I., Navas, E., Rodríguez, J. C., Gándara, R. (1999) "Improving quality in a speech synthesizer based on the MBROLA algorithm." Proc. Sixth European Conference on Speech Communication and Technology, pp. 2299-2302, Budapest. ISSN: 1018-4074

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

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

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

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

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

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.