The Lessac Technologies Time Domain Diphone Parametric Synthesis System for Microcontrollers for Blizzard Challenge 2012

Mike Baumgartner, Reiner Wilhelms-Tricarico, John Reichenbach

{mike.baumgartner, reiner.wilhelms, john.reichenbach} @lessactech.com

Abstract

Advances in the capabilities of microcomputer systems have opened the door to new approaches to real time speech synthesis. In the past, diphone synthesis was a popular synthesis method. More recently, unit selection speech synthesis has afforded higher quality synthesis, mainly by eliminating the need for significant signal processing, and thus preventing the signal processing artifacts that are the consequences of speech segment modifications. Instead, unit selection synthesis consists substantially of real segments of unaltered speech. It was hoped that with large enough voice databases, that could provide enough recorded sections of speech, there would be sufficient coverage for any utterance required for speech synthesis. Even as unit selection speech synthesis system databases have become considerably larger, the realization of constructing natural speech entirely from segments of unaltered speech units has still fallen short of expectations. The Blizzard Challenge has provided a measure to quantify how much of a difference in quality has transpired in the new unit selection approaches compared to the old diphone synthesis methods. This diphone synthesis system also is an example of working towards a goal of high quality synthesis that still works on very limited hardware resources.

Index Terms: Speech Synthesis, Blizzard Challenge, Diphone

1. Introduction

Initially, our intent was to explore whether advances in front-end labeling systems, such as the Lesseme system used by Lessac Technologies in its unit selection system could also be used to drive a very small compact parametric synthesizer suitable for running on ten dollar microcontrollers. We explored multiple parametric synthesis approaches, but mainly for reasons of time and available resources, we eventually defaulted to building a demonstration system based on the many years of diphone work.

Why enter an older technology in the Blizzard Challenge? Because this time-domain diphone synthesis was designed for low cost microcontroller applications, and was expected to provide usable quality. As we learned from the listening test results, this diphone system is less natural than current unit selection systems. However, while among the lower ranking systems, this diphone synthesizer is ranked just slightly lower than most of the conventional unit selection systems when evaluated on an MOS basis for sentences. For the longer paragraph length sections of synthesis, this time domain diphone synthesis does not measure up as well.

This is a what-if scenario. If used in a low cost system with limited resources, what difference in quality can be expected when compared to the best unit selection systems currently available? Blizzard Challenge 2012 is not an ideal structure for evaluating time domain diphone parametric synthesis. The John Greenman Librivox voice is not a phonetically balanced corpora. The voice corpus was recorded in relatively poor conditions with a fair amount of background noise. The mp3 lossy compression of the original source for the voice corpus introduced difficulties in pitch-marking and signal discrimination. Despite these hurdles, we were successful in building a moderate quality voice with comparatively few, but still highly noticeable, artifacts. The results of the Blizzard Challenge were helpful in gauging what could be expected.

2. Voice Building

There were some technical issues with selecting very small segments of speech from the several thousand prompts that were suitable for building a unit selection voice. A diphone database generally consists of one copy of small segments of speech. Instead of trying to output the phones and joining between phones, the diphone system joins in theory at the middle of the phone where the spectra are stable (Olive et. al. 1998). Therefore all combinations of phone to phone segments of speech that can occur need to be included in the database.

This system started with the Lessemes used in the Lessac unit selection voice and converted them into 46 phones. If all possible combinations occurred in the English language, there would be 46x46 diphones or 2,116 diphones needed. In practice, the figure is more around 1,500. The challenge is to find the best example of each diphone that is most representative of that diphone, that will match in duration, pitch and spectra without any signal processing, and that is most likely to join smoothly with other diphones in numerous linguistic contexts.

Using available unit selection toolsets, one automatic way to achieve this would be to build a unit selection voice with all the data. Next, limit the selection of speech segments to diphones only. Then, run the unit selection synthesis using large text corpora that will cover all diphone combinations in a statistically large enough sample. A diphone use log could be amassed and then from the statistics of use, a diphone selection could be made. Because of the unit selection processes this might lead to multiple diphone paths that might not converge to a single best choice per diphone.

The voice building approach for this synthesis for the Blizzard Challenge was done in a very simple fashion. A set of programs was written to step through the phone label files and build up a diphone map. We loaded the map and either selected the diphones by hand, or defaulted to the first diphone on the map. The selection program allowed testing the sound of the diphone when fitted to others. The more frequently occurring diphones were selected by hand, and absent noting particularly bad results, the less frequently occurring diphones defaulted to the first one that occurred in the map. Because of time constraints the first one in the map was used fairly often. One might assume that with such a large database that the distribution of the number of each diphones available would generally be more or less even. This was not the case.

After the diphone selections were made, the diphone database was assembled and tested by synthesis. Areas that did not have suitable sounding phone targets were noted and better diphones were substituted. Unlike with unit selection, since the same small set of diphone segments of speech are used over and over again, if they are poorly chosen, there is a very noticeable degradation in synthesis quality.

The goal was also to limit selections that involved substantial phonetic co-articulation effects. Diphones from such text locations found in the acoustic data sound good in some linguistic contexts, but quite out of place, and mismatched in others.

The favorable thing with a diphone voice is that changes can be incorporated almost instantly using a current Pentium class system to build the voice database.

3. Text to Phone, Prosaic and Duration Section

One of the goals of this entry was to limit differences in quality in the parametric parameters driving this diphone synthesis. We hoped to use an unchanged Lessac front-end to provide the parameters to drive the diphone back-end synthesizer. Then only the synthesis section would be different when compared to other systems.

One of the potential target applications for this system is low cost audio book synthesis. Instead of the thousands of bytes per second required for speech compression systems, tens of bytes per second would be required for phones and pitch information. Then instead of one book per device, a library of books could be held. In other words, most target applications would not require having a full text to speech system included. Only the synthesis section would be included.

It was hoped that the parameters that drive the large Lessac system could be used just the same. It was discovered however, the pitch targets that drive the larger Lessac unit selection system were not sufficient to provide a complete high quality set of parameters for diphone synthesis. Unit selection does not change the localized phone pitch in speech segments. Since unit selection uses larger sections of real speech, the natural localized phone pitch variations are automatically included in the final synthesis. This is one of the good qualities of unit selection. This is not the case with diphone synthesis. Without the minor localized pitch variations inherent in human speech, our initial diphone synthesizer sounded quite flat and robotic.

To include localized pitch variations in this parametric synthesis, a localized variation pitch map was constructed using the Nancy voice from last year's Blizzard Challenge. The entire database phone map was constructed with the pitch variations. Then using a dynamic programming like technique, the largest sections of phone sequences that matched the phone sequence of the required utterance were built up. Then these localized variations were added to the pitch targets. In essence, the Nancy voice was driving the subtle prose elements of the parametric synthesis. It was noted that some of the personality of the Nancy voice could be recognized in the synthesis. The down side to this technique was that sections that had the same phone sequences in close proximity tended to have unnatural pitch repetitions. This could have been avoided by adding code that insured different areas of the localized phone variation map were used for each utterance chunk.

4. Synthesis Section

Like the Lessac unit selection synthesis, the diphones are concatenated together in a process that works entirely in the time domain. The concatenation of voiced sounds is done pitch synchronously, and some mutual adjustments of two sounds that are concatenated are made to increase the coherence and to reduce clicks and warbles. The overall pitch and duration changes are added to match the parameters given.

There are limits to the quality of a diphone system. The better you can discern mismatches by ear, the better synthesis you will have. With the approach we used for building a parametric diphone synthesizer, diphone selection was entirely based on listening and accepting or rejecting individual diphones. Therefore like the Lessac unit selection, the choices of the units or diphones used, and not the synthesis and concatenation technique itself makes for better synthesis. Unlike the larger unit selection systems, where the units to be concatenated are chosen in real-time based on join and target costs, in this diphone parametric system these diphone unit choices are made ahead of time when building the voice, and at maximum only one unit exists in this system for each actually appearing diphone. Had we had the time to more carefully and optimally select the diphones for the voice, one could expect a modest improvement in quality.

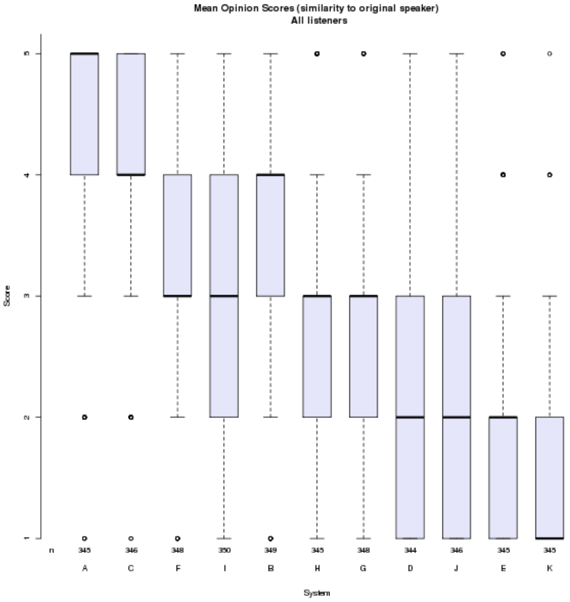
5. Results

Ten systems participated in the Blizzard Challenge 2012. Natural, as recorded, human speech was also evaluated as an eleventh pseudo system (system A). One of the systems (system B) was a benchmark system to allow approximate comparison with previous Blizzard Challenges. This benchmark system was built using Festival by CSTR in a manner similar to their 2007 entry. The large Lessac hybrid concatenation unit selection system was system F. The Lessac time domain diphone parametric synthesis system that is discussed in this paper is system J.

During the online listening evaluation, listeners were asked to judge samples of synthesized speech for both naturalness, and similarity to the original speaker. Both of these tasks were rated by listeners on a mean opinion scale (MOS). Each listener was also asked to listen to synthesized semantically unpredictable (nonsense) sentences (SUS), and transcribe the words that they heard.

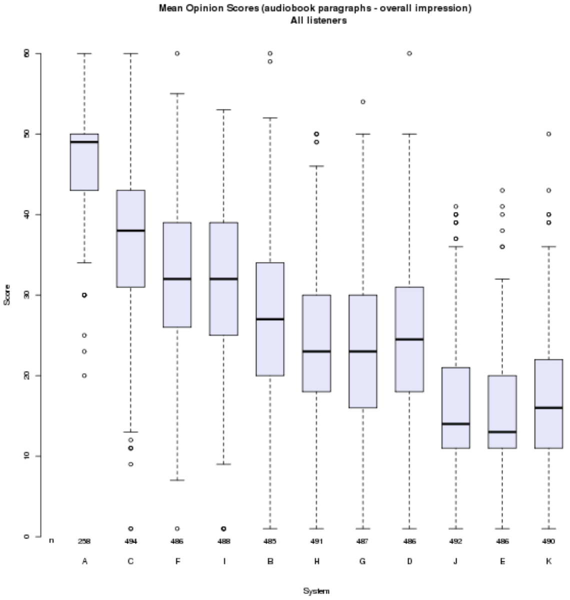
Naturalness and similarity to original speaker – sentences

A mean opinion scale (MOS) ranging from 1 to 5 was used to evaluate how both how natural synthesized speech is sounds, as well as how similar it is to natural human speech. This is the same assessment method that has been used in previous Blizzard Challenges. With respect to similarity to the original speaker, the Lessac diphone parametric synthesizer (system J) received an MOS score of 2.4. Of the ten systems overall, six systems ranked higher, and three systems ranked lower. For naturalness, the MOS score was 1.9.



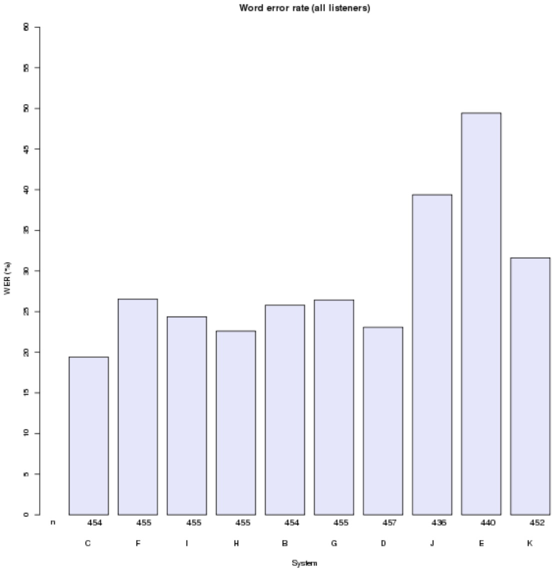
Naturalness and similarity to original speaker – paragraphs

A mean opinion scale (MOS) ranging from 1 to 60 was used to evaluate pleasantness, naturalness, speech pauses, stress, intonation, emotion and listening effort. On an overall basis, the Lessac diphone parametric synthesizer (system J) received an MOS score of 16.



Word Error Rate

For the semantically unpredictable sentences (SUS), the Lessac diphone parametric synthesizer (system J) received a mean word error rate of 39%, among the worst word error rates of the systems being evaluated...



6. Conclusion

In the past, companies would often choose voice talents based on specific voice characteristics that were thought to synthesize better, or with fewer artifacts. For example, many TTS voices have sharp vocal chords to help mask artifacts that would otherwise occur in the synthesis. By choosing voice talents that already have these attributes in their original voice; it is less anomalous when the synthesizer produces more of these attributes as a result of the synthesis process.

For the Blizzard Challenge 2012, each participant built a John Greenman voice. Given that this voice had been reconstructed from mp3 lossy digital compression, and was not selected for voice characteristics that were likely to synthesize well, it may not have been an ideal voice for building a diphone synthesizer. From the listening test plot results on the previous page, it is obvious that our diphone system was outclassed in this competition. However, despite the poor showing, when compared to synthesizers of just a few years ago, this compact, low overhead diphone synthesizer might compare quite favorably.

Despite the poor ranking, many goals were achieved. We demonstrated that the prosody characteristics of another voice can be used to drive and enhance parametric synthesis. While doing this enhanced the perceived naturalness of the synthesized voice, it probably negatively impacted the similarity scores as compared with the original voice talent. We demonstrated that a diphone voice can be extracted from large voice corpora in a brief period of time. We also demonstrated that many of the tools and techniques that have been developed for unit selection approaches to voice synthesis can be used in building diphone voices. We clearly demonstrated that a diphone voice can be built without the buzziness of LPC synthesis.

We hope to be able to demonstrate that by choosing a voice model that better meets the needs of diphone synthesis, the perceived quality will improve. We hope that further development of the diphone approach, or other parametric approaches, to synthesis will allow us to approach the quality of unit selection systems, while retaining the very small size and footprint of this diphone parametric system.

The perceived synthesis quality, while lower than that of unit selection systems, may show some promise for certain target applications. Though not a mainstream product, it would be nice to have a library of audio books that could be heard on a ten dollar device. The buttons on such a device could have voice prompts so if someone was sight impaired or driving a car, they could make selections without looking. If the synthesis footprint is compact, and based on phones and pitch information only, toys such as those given away at fast food restaurants could have a small audio book included as part of the package, all in the not too distant future. This is becoming more possible as memory densities grow in low cost devices. It appears that diphone synthesis might still be a usable speech synthesis technique.