WHICH IS MORE IMPORTANT IN A CONCATENATIVE TEXT TO SPEECH SYSTEM—PITCH, DURATION, OR SPECTRAL DISCONTINUITY?

M. Plumpe, S. Meredith
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA

ABSTRACT
This paper focuses on experimental evaluations designed to determine the relative quality of the components of the Whistler TTS engine. Eight different systems were compared pairwise to determine a rank ordering as well as a measure of the quality difference between the systems. The most interesting aspect of the results is that the simple unit duration scheme used in Whistler was found to be very good, both when it was used in combination with natural acoustics and pitch as well as when it was taken in combination with synthetic pitch. The synthetic pitch was found to be the aspect of the system that results in greatest quality degradation.

1. INTRODUCTION
We have presented Whistler, Microsoft’s Trainable Text-To-Speech (TTS) system in [1][2]. We will primarily look at three aspects of the system, the pitch (fundamental frequency), phoneme duration, and acoustics. Whistler has a concatenative synthesizer, using context-dependent phoneme units that are automatically selected from a training database. The pitch is generated by rule, while the durations are generally the average duration for the unit of interest.

In order to learn what area needed the most research attention, we ran a study comparing eight different versions of the Whistler TTS system. All versions are from the same speaker. One version is original speech, the other versions have one or more components of Whistler added to “degrade” the signal. The final version is our complete TTS system.

While the particular results of this study are clearly dependent on the Whistler Speech Synthesizer, it is somewhat unique amongst previous studies in that it attempts to identify more closely the cause of quality degradation. Also, as previous internal studies have shown the Whistler engine to be of similar quality to other commercially available speech synthesizers, we hope that to some extent these results can be extended to other concatenative synthesizers.

In section two we discuss Whistler. Section three describes the experimental setup. Section four discusses the results of the study, while section five has conclusions.

2. WHISTLER
We will now briefly describe the Whistler TTS engine in order to better understand the systems being compared in this evaluation. A block diagram of the engine is shown in Figure 1.

Figure 1: Block diagram of the Whistler synthesizer. Either the front end can provide the phonemes, pitch targets, and phoneme durations, or any of these can be provided through transplanted prosody. The unit inventory can be augmented with additional units to allow the original acoustics to be used.

Whistler can use either natural or synthetic pitch, phoneme durations, amplitudes, and speech units, and the speech units can optionally be compressed. Some quality degradation will occur even when the natural versions of all the components are used. This is due to any non-perfect reconstruction characteristics. Primary among these is that pitch and amplitude can only be specified three times per phoneme, and the output pitch and amplitude is linearly interpolated between these specified values. These values occur at the beginning, middle, and end of each phoneme. This brings in several areas of quality degradation, namely quality loss due to prosody modification as well as a reduction in naturalness due to lost microprosody.
While in general we have attempted to separate out each aspect of the engine to determine its impact independently, these degradations are always present.

For the conversion of text to phonemes and generation of pitch contours, we have used a text analysis component derived from Lernout & Hauspie’s commercial TTS system [5]. This component performs text analysis functions, including text normalization, grapheme-to-phoneme conversion, part-of-speech tagging, shallow grammatical analysis, and prosodic contour generation. Alternatively, phonemes and pitch targets can be extracted from natural speech and provided to the engine using a transplanted prosody format.

Aside from the pitch and phonemes, all remaining aspects are dependent on the training, which we will now describe. Whistler uses decision-tree clustered phone-based units [1]. Each unit is a cluster of phones, whose phonetic contexts and other characteristics, such as stress, are used to traverse automatically trained decision trees for finding the cluster. Unit duration is not taken into account in our current decision trees. The trees are trained from a large speaker independent database. One benefit of using a decision tree is that any number of nodes can be selected, the system used in this evaluation has approximately 3000 units.

The unit acoustics and durations are determined from a single speaker training database. The database consists of approximately 7 hours of speech collected as isolated sentences. There are approximately 100,000 individual phoneme instances in the database, giving on average 30 instances of each phoneme unit. The database was constructed to attempt to have at least 10 instances of each unit, while extremely common units may occur hundreds of times.

One actual instance of speech must be selected to represent each unit. In order to select this unit, a single speaker training database is segmented using the Whisper speech recognition engine [3]. The segmentation provides a score for each unit through the probability of the HMM evaluated for segmentation. After discarding unit instances whose pitch, duration, or amplitude are outliers, the instance with the highest HMM probability is retained to represent this unit in the synthesis database. Since we have some measure indicating that each unit itself is of good quality, quality loss from the synthetic units is primarily due to mismatch at the concatenation point or degradation due to the prosody modification algorithms. The units are compressed to limit the size of the total system for practical purposes.

The synthetic durations used are in general the mean duration, determined by the segmentation above, of all instances of the appropriate unit seen during training (outliers are not discarded). The engine has one very simple rule to extend the duration of units before a silence. The syllable coda before a sentence ending is lengthened 30%, while the syllable coda preceding other pauses is lengthened 20%. If natural durations are desired instead of the synthetic values, they can be extracted from natural speech and provided to the engine through the transplanted prosody format.

The average amplitude of each unit is similarly made equal to the average amplitude of all instances of that unit seen in training, amplitude is generally not specified at synthesis time. Amplitude was mostly ignored in this study. Natural amplitude are used in systems A and B (systems are described below), while all other systems used the default average amplitudes.

3. EXPERIMENTS

We will now describe the experimental evaluation. The experiment was run by an individual not familiar with the systems being evaluated. In this way, we hope to avoid any influence on the results based on our preconceived opinions.

There are eight possible combinations of the three aspects of the synthesizer we wish to look at, acoustics, pitch, and phoneme durations. Of these, we look at all but two. At one end is the system comprising natural versions of all the components. The duration of each phoneme is determined by forced alignment using the Whisper speech recognizer. We have previously found that only 4% of sentences contain a segmentation that differs by more than 20ms from hand labeled segmentation [1], thus we can consider the determined durations to be sufficiently accurate. The pitch is extracted from natural speech for each pitch period using a laryngograph. The three values passed to the synthesizer are the endpoints of lines determined through a minimum mean squared error estimation of both lines simultaneously.

In order to use natural units (acoustics), we break the sentence up into phonemes based on the segmentation given by Whisper. These units are compressed and appended to the unit inventory, and the engine is instructed to use these natural units instead of the default units. As it is possible that the natural speech contains a different pronunciation of a word than the synthesizer would produce (due to dialect, or other choice in word pronunciation), the phonemes are also given by Whisper and these passed to the engine as well. We call this system “framework,” because the natural version of each variable is used, and any quality degradation occurs due to the framework of the synthesizer.

At the other end of the spectrum is the all synthetic system. The phonemes and pitch are determined by rule as described in the previous section. The phoneme durations are determined by table lookup and the one rule as described above. The acoustic units are those from the default unit database, selected by the procedure described in the previous section.

Of the other six possible combinations, all but natural pitch with synthetic durations and units; and natural durations with synthetic units and pitch were included in the trial. A complete listing of the systems included in the evaluation is shown in Table 1. In addition to these six combinations, two other systems were added. First are the recorded natural utterances, the baseline. Second are the natural utterances that have been compressed. This system was included as a second baseline, as all other systems are compressed.
In order to fully compare all the systems without any a priori knowledge of quality ranking, each system was compared to every other system. For each utterance, the subjects heard the utterance from two systems, separated by a one-second pause, and indicated which utterance they preferred. The subjects did not have the option of choosing no preference. Reaction times were measured to help determine the ease of distinguishing quality between two systems. The reaction time was also used to verify that the subjects were waiting until both utterances had completed before making up their mind. Therefore, reaction times under 250 ms resulted in that trial being discarded. The systems were counterbalanced for order to ensure that the ordering of the systems did not influence the results.

Fifteen utterances were used, the same utterances for all systems. The utterances, all sentences, are listed in the appendix. The utterances vary from under one second to almost 15 seconds, with an average of slightly over four seconds. For all systems, a 22 kHz sampling rate was used. The utterances were chosen for their representation of a variety of prosodic situations as well as phonemic coverage. All utterances were based on the same speaker, the female speaker used to train Whistler’s female voice. The 15 utterances were recorded by this speaker, with the variables of interest either extracted from these natural utterances or generated by the synthesizer.

Twenty subjects participated in the experiment, with approximately equal numbers of males and females. All subjects were screened for hearing impairment. The subjects were briefed as to the purpose of the experiment. They were instructed to indicate their preferences for the utterances based on how natural the utterances sounded, so that utterances that sound more like natural human speech would be preferred. The subjects wore headphones to enable multiple subjects to be run simultaneously. Our experience indicates that this tends to accentuate errors in compression and acoustics. We have found informally that the compression is approximately transparent for speech played through standard PC multimedia speakers.

With eight systems, each compared against all others as well as itself, a total of 36 system versus system comparisons are needed, as shown in Table 2. With 15 utterances per system, this gives 540 comparisons of two utterances each. In order to help counteract fatigue effects, each subject listened to half of the comparisons, resulting in 270 comparisons per subject, taking about two hours. All subjects heard all 15 utterances and all 36 comparisons, but only a subset of the combinations. The 270 comparisons were divided into three blocks of 90 trials to allow for breaks.

In order to fully compare all the systems without any a priori knowledge of quality ranking, each system was compared to every other system. For each utterance, the subjects heard the utterance from two systems, separated by a one-second pause, and indicated which utterance they preferred. The subjects did not have the option of choosing no preference. Reaction times were measured to help determine the ease of distinguishing quality between two systems. The reaction time was also used to verify that the subjects were waiting until both utterances had completed before making up their mind. Therefore, reaction times under 250 ms resulted in that trial being discarded. The systems were counterbalanced for order to ensure that the ordering of the systems did not influence the results.

Fifteen utterances were used, the same utterances for all systems. The utterances, all sentences, are listed in the appendix. The utterances vary from under one second to almost 15 seconds, with an average of slightly over four seconds. For all systems, a 22 kHz sampling rate was used. The utterances were chosen for their representation of a variety of prosodic situations as well as phonemic coverage. All utterances were based on the same speaker, the female speaker used to train Whistler’s female voice. The 15 utterances were recorded by this speaker, with the variables of interest either extracted from these natural utterances or generated by the synthesizer.

Twenty subjects participated in the experiment, with approximately equal numbers of males and females. All subjects were screened for hearing impairment. The subjects were briefed as to the purpose of the experiment. They were instructed to indicate their preferences for the utterances based on how natural the utterances sounded, so that utterances that sound more like natural human speech would be preferred. The subjects wore headphones to enable multiple subjects to be run simultaneously. Our experience indicates that this tends to accentuate errors in compression and acoustics. We have found informally that the compression is approximately transparent for speech played through standard PC multimedia speakers.

With eight systems, each compared against all others as well as itself, a total of 36 system versus system comparisons are needed, as shown in Table 2. With 15 utterances per system, this gives 540 comparisons of two utterances each. In order to help counteract fatigue effects, each subject listened to half of the comparisons, resulting in 270 comparisons per subject, taking about two hours. All subjects heard all 15 utterances and all 36 comparisons, but only a subset of the combinations. The 270 comparisons were divided into three blocks of 90 trials to allow for breaks.

In order to fully compare all the systems without any a priori knowledge of quality ranking, each system was compared to every other system. For each utterance, the subjects heard the utterance from two systems, separated by a one-second pause, and indicated which utterance they preferred. The subjects did not have the option of choosing no preference. Reaction times were measured to help determine the ease of distinguishing quality between two systems. The reaction time was also used to verify that the subjects were waiting until both utterances had completed before making up their mind. Therefore, reaction times under 250 ms resulted in that trial being discarded. The systems were counterbalanced for order to ensure that the ordering of the systems did not influence the results.

Fifteen utterances were used, the same utterances for all systems. The utterances, all sentences, are listed in the appendix. The utterances vary from under one second to almost 15 seconds, with an average of slightly over four seconds. For all systems, a 22 kHz sampling rate was used. The utterances were chosen for their representation of a variety of prosodic situations as well as phonemic coverage. All utterances were based on the same speaker, the female speaker used to train Whistler’s female voice. The 15 utterances were recorded by this speaker, with the variables of interest either extracted from these natural utterances or generated by the synthesizer.

Twenty subjects participated in the experiment, with approximately equal numbers of males and females. All subjects were screened for hearing impairment. The subjects were briefed as to the purpose of the experiment. They were instructed to indicate their preferences for the utterances based on how natural the utterances sounded, so that utterances that sound more like natural human speech would be preferred. The subjects wore headphones to enable multiple subjects to be run simultaneously. Our experience indicates that this tends to accentuate errors in compression and acoustics. We have found informally that the compression is approximately transparent for speech played through standard PC multimedia speakers.

With eight systems, each compared against all others as well as itself, a total of 36 system versus system comparisons are needed, as shown in Table 2. With 15 utterances per system, this gives 540 comparisons of two utterances each. In order to help counteract fatigue effects, each subject listened to half of the comparisons, resulting in 270 comparisons per subject, taking about two hours. All subjects heard all 15 utterances and all 36 comparisons, but only a subset of the combinations. The 270 comparisons were divided into three blocks of 90 trials to allow for breaks.

## Table 1: The various systems evaluated are listed here, along with a label to simplify references.

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Natural sentences</td>
</tr>
<tr>
<td>B</td>
<td>Compressed natural sentences</td>
</tr>
<tr>
<td>C</td>
<td>Framework</td>
</tr>
<tr>
<td>D</td>
<td>Natural units and pitch, synthetic durations</td>
</tr>
<tr>
<td>E</td>
<td>Synthetic units, natural pitch and duration</td>
</tr>
<tr>
<td>F</td>
<td>Natural units and durations, synthetic pitch</td>
</tr>
<tr>
<td>G</td>
<td>Natural units, synthetic pitch and durations</td>
</tr>
<tr>
<td>H</td>
<td>Synthetic units, pitch, and durations</td>
</tr>
</tbody>
</table>

After the final block of trials, participants completed an additional measure of preference for the eight systems. Participants listened to all eight systems’ versions of utterance number 9, rating each on a scale of 1 – 10, where 1 meant “it sounds awful” and 10 meant “it sounds perfect.” This task was administered to corroborate results from the primary measure. It is similar to the Mean Opinion Score tests used in speech coding. The eight versions of utterance number nine are included on the CD-ROM version of the proceedings.

### 4. RESULTS

To convert from the 36 preference tests to an ordered ranking of the eight systems, the total number of times a system was preferred was divided by the total number of times that system appeared in a trial (excluding self-comparisons), giving a preference percentage. The systems were then ranked by these percentages. In order to measure the significance of this ranking, preference percentages were calculated for each subject, then a repeated-measures ANOVA was performed to measure the significance of the rankings.

The results for the primary measure are given in Tables 3. As is expected, the top three systems are the natural, compressed, and framework systems. One key finding is that the synthetic durations, despite their simplicity, are very good. In comparing the systems with natural and synthetic durations, it is apparent that the largest degradation occurs at phrase endings and more complex syntactic structures, such as lists. For short sentences, it is nearly impossible to distinguish between synthetic and natural durations.

When synthetic durations were used along with synthetic pitch, the evaluation shows a minimal distinction from just using synthetic pitch alone. This indicates that the method used to estimate durations is likely sufficient for many different systems.
The secondary measure yielded similar results, as shown in Table 4. The results are of less statistical significance because they came from one-fifteenth the data.

<table>
<thead>
<tr>
<th>System</th>
<th>Preference Percentage</th>
<th>Statistical Significance of change in Preference Percentage</th>
<th>95% Confidence Interval for preference percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>.91</td>
<td>A &gt; B: F(1,19) = 31.808, p &lt; .001</td>
<td>.886 &lt; .91 &lt; .934</td>
</tr>
<tr>
<td>B</td>
<td>.79</td>
<td>A &gt; B: F(1,19) = 31.808, p &lt; .001</td>
<td>.762 &lt; .79 &lt; .818</td>
</tr>
<tr>
<td>C</td>
<td>.68</td>
<td>B &gt; C: F(1,19) = 33.413, p &lt; .001</td>
<td>.654 &lt; .68 &lt; .706</td>
</tr>
<tr>
<td>D</td>
<td>.54</td>
<td>C &gt; D: F(1,19) = 22.041, p &lt; .001</td>
<td>.499 &lt; .54 &lt; .581</td>
</tr>
<tr>
<td>E</td>
<td>.41</td>
<td>D &gt; E: F(1,19) = 6.905, p = .17</td>
<td>.347 &lt; .41 &lt; .473</td>
</tr>
<tr>
<td>F</td>
<td>.27</td>
<td>E &gt; F: F(1,19) = 25.873, p &lt; .001</td>
<td>.233 &lt; .27 &lt; .307</td>
</tr>
<tr>
<td>G</td>
<td>.23</td>
<td>F &gt; G: F(1,19) = 3.560, p = .075</td>
<td>.205 &lt; .23 &lt; .255</td>
</tr>
<tr>
<td>H</td>
<td>.17</td>
<td>G &gt; H: F(1,19) = 10.628, p = .004</td>
<td>.139 &lt; .17 &lt; .201</td>
</tr>
</tbody>
</table>

Table 3: Shown here are the rankings by preference percentage, along with the statistical significance from the ANOVA analysis and 95% confidence intervals. For the statistical significance column, each system is compared to the system that ranked one above it. Shown is the F statistic from ANOVA and the corresponding probability (p) that the difference in preference percentage isn’t significant. For four of the rankings there is a less than 0.1% chance that the order isn’t significant, the others are as shown. As can be seen, all rankings except the F to G ranking are statistically significant.

The reaction times in general agreed with the ANOVA analysis for the significance of the difference between systems and the rankings. For example, in comparing to system A, natural speech, the average response time for system B was 875 ms, while for system G it was 575 ms. Thus it on average it took the subjects more time to determine a preference when the difference was minor.

In Figure 2 we plot the ratings from the two measures. This figure illustrates that systems F, with synthetic pitch, and G, with synthetic pitch and duration, are nearly equal in quality. The secondary measure indicates very little difference between systems B, C, and D, indicating that passing the speech through the framework of the synthesizer and using synthetic duration results in minimal quality loss.

The reaction times in general agreed with the ANOVA analysis for the significance of the difference between systems and the rankings. For example, in comparing to system A, natural speech, the average response time for system B was 875 ms, while for system G it was 575 ms. Thus it on average it took the subjects more time to determine a preference when the difference was minor.

5. CONCLUSIONS

This study confirmed our initial hypothesis that the pitch generation component of the Whistler TTS engine is the component that has the largest impact on quality degradation. The fact that the synthetic durations reduced quality only minimally in two situations, with natural and synthetic pitch, indicates that the simple clustering method used to determine the average duration does an excellent job.

By no means were all interesting systems studied. Further areas of interest include the impact of using headphones versus speakers, removing compression, and verifying the assumption that amplitude is of lesser importance than duration and pitch.
6. ACKNOWLEDGEMENTS

The authors would like to express their gratitude to Scott Tiernan and Mary Czerwinski for their help in designing and running this study.

A1. SENTENCES

We now list the 15 sentences used in the study.

1. Have you come to any conclusion?
2. I wonder, by the way, who will be named director?
3. Several delegates, he among them, will state their opposition at the next meeting.
4. Who washed the car?
5. We hold these truths to be self-evident; that all men are created equal; that they are endowed by their creator with certain inalienable rights; that among these are life, liberty, and the pursuit of happiness.
6. Who said, “It ain’t over till it’s over?”
7. The due date, once the loan has been approved, can be the date most convenient for you.
8. Look at that!
10. Strictly between the two of us, do you think she’s crazy?
12. And in science fiction, tiny computers recognize speech, understand it, and even reply.
13. How much will it cost to do any necessary modernizing and redecorating?
14. How much and how many profits could a majority take out of the losses of a few?
15. Will you please confirm government policy regarding waste removal?

7. REFERENCES


