

# A Component by Component Listening Test Analysis of the IBM Trainable Speech Synthesis System

*Robert E. Donovan*

IBM T.J.Watson Research Center, PO Box 218  
Yorktown Heights, New York, 10598, USA

red@watson.ibm.com

## Abstract

This paper reports on a listening test conducted to determine the impact on speech quality of each component in the IBM Trainable Speech Synthesiser. The study was originally conceived to direct future research effort to those components with the greatest potential for improvement. However, the results and conclusions regarding prosodic modification, concatenation unit length, and decision tree clustering are generally applicable and may be of wider interest.

## 1. Introduction

Synthetic speech generated by a Text-to-Speech (TTS) system is usually of a lower quality than the equivalent natural speech. This inferiority is the combined result of many system components working less than perfectly. For a modern concatenative speech synthesis system these system components might include text normalisation, prediction of phrase break locations, text to phone conversion, prosody generation, segment selection, prosodic modification and segment combination. Although users may be aware of specific failings of individual components of a system from time to time, it is usually not obvious how much effect the failing of each component has on overall system performance.

This paper presents the results of a listening test conducted to determine, for the IBM Trainable Speech Synthesis System described in [1], how much effect many of the system components mentioned above have on overall system performance. The test was conducted using speech generated from eight different systems ranging from natural speech to (almost) pure TTS synthesis, with each system introducing one more synthesis component than the last. The test was conducted principally to help direct future research effort to those components with the most room for improvement.

The current IBM synthesiser operates as described in [1] minus the Backing-Off algorithm and plus an independent rule-based front-end which performs text normalisation, phrase break placement, text to phone conversion and pitch and duration prediction. Familiarity with [1] will greatly aid understanding of the current paper.

## 2. Test Stimuli

The eight different systems used to supply speech (at an 8kHz sampling rate) for use in the listening test are described in the following sections.

Modification	75th Percentile
Duration Increase	1.366
Duration Decrease	0.536
Energy Increase	1.960
Energy Decrease	0.287
Pitch Increase	1.072
Pitch Decrease	0.929

Table 1: 75th percentiles of the modification-factor distributions seen in synthesis for different types of prosodic modification.

### 2.1. System 1

Ten news-wire style sentences, selected at random, were synthesised using the IBM synthesiser. As described in [1], during synthesis the system attempts to select segments from the synthesis database which have, as far as possible, prosody close to the target values required. The selected segments are then modified using signal processing to have the target prosody. Table 1 tabulates the difference between the selected segments' prosody and the corresponding target prosody for these ten sentences. It shows, for each type of prosodic modification, the 75th percentile of the distribution of modification factors.

Another ten news-wire style sentences were selected at random and recorded by the same, male, native American English speaker used to train the synthesis system. The speech was recorded using the same equipment and setup as the speech used to build the synthesis system. Each sentence was read ten times, initially five times with very different prosody, and subsequently five times with very similar prosody.

The recorded sentences were processed using the same procedure used in synthesis system construction [1]. In summary, a phonetic transcription was obtained by aligning pronunciations from a dictionary to the speech using (mostly) 3-state left-to-right speaker-dependent cross-word decision-tree state-clustered hidden Markov models (HMMs). Where multiple pronunciations were available the selection between alternatives was made by the HMMs. The HMMs were then used to produce a state level alignment of the speech defining the segments to be used in synthesis. A laryngograph signal recorded in stereo with the speech was used to determine the moments of glottal closure and hence the fundamental frequency.

The objective in recording multiple versions of each sentence was to obtain at least three versions of each sentence with identical phonetic transcriptions whose prosody differed from each other by amounts similar to those listed in Table 1. The sentences recorded with very different prosody turned out to have much more variation than required. However, the sen-

tences recorded with similar prosody turned out to differ by approximately the required amount, see Section 2.3. Of the 10 sentences 8 had at least 3 readings with identical phonetic transcriptions. Of these three one was chosen to be the System 1 speech, the natural speech for the listening test.

The synthetic energy target values used in Systems 7&8 are obtained from the medians of the leaves of the energy prediction trees estimated from the database during synthesis system construction, [1]. Since they are based on median energies in appropriate contexts these predicted energy values ensure that a large number of database segments in each context are able to synthesise the predicted energy. In order to obtain the highest speech quality from Systems 5&6 the energy values extracted from the System 1 speech should match the database segments in a similar way. The volume level of the System 1 speech was therefore set such that the 90th percentile of its energy values was 1.3dB quieter than the 90th percentile of the energy-tree predictions for the same sentences, and the median of its energy values 1.0dB louder than the median of the energy-tree predictions.

## 2.2. System 2

System 2 speech introduces the minimal amount of signal processing degradation the synthesis system can introduce. The speech was produced by using the pitch, duration, and energy values extracted from the System 1 speech as prosodic targets during synthesis. The synthesiser’s runtime database was augmented with the segments of the System 1 speech, and the cost function reward for using contiguous speech made very large. The result was that all the System 1 segments were concatenated to produce the System 1 prosodic targets. The speech generated therefore uses contiguous segments throughout from ideal contexts and has ideal (ie. natural) prosodic targets which match the prosody of the segments producing them. The synthesiser in this case is essentially taking apart an utterance and putting it back together. Ideally this process would be lossless, but in practice some degradation creeps in. Pitch targets are specified at the beginning and end of each state and linearly interpolated inbetween during synthesis. Thus, in mid-state a small amount of prosodic modification may be occurring, exercising IBM’s (unpublished) prosody modification algorithms.

## 2.3. System 3

System 3 speech introduces prosodic modification by amounts similar to those typically seen in synthesis. The speech was produced in exactly the same way as the System 2 speech, except that the runtime database was augmented with the segments of one of the *other* readings of each sentence. The speech generated therefore uses contiguous segments throughout from ideal contexts and has ideal prosodic targets, but must modify the prosody of the segments used. The amounts of prosodic modification introduced are shown in Table 2, and should be compared to the values in Table 1.

## 2.4. System 4

System 4 speech introduces segment concatenation points to approximately the same extent that they occur in synthesis. Analysis of the 10 pure synthetic sentences mentioned in Section 2.1 shows that concatenation typically occurs on 70% of state boundaries during synthesis. System 4 speech was prepared in exactly the same manner as System 3 speech, except that the runtime database was augmented with the segments

Modification	75th Percentile
Duration Increase	1.34783
Duration Decrease	0.553191
Energy Increase	2.20967
Energy Decrease	0.494295
Pitch Increase	1.06086
Pitch Decrease	0.940961

Table 2: 75th percentiles of the modification-factor distributions seen in System 3 speech for different types of modification.

of *two* other readings of each sentence. The synthesiser cost function was modified to select only from these augmented segments, and the reward for using contiguous segments changed to a small penalty. The result was that the system spliced segments from only the two other readings to generate the speech. The penalty for using contiguous segments was adjusted to obtain concatenation points on 76% of state boundaries. The speech generated therefore uses segments from ideal contexts and has ideal prosodic targets, but uses non-contiguous segments and must modify the prosody of the segments used.

## 2.5. System 5

System 5 speech introduces the use of decision trees and spectral continuity costing to identify segments to use in synthesis, instead of using segments from human readings of the target sentence as in Systems 1-4. The construction of the decision trees and the use of the continuity cost are described in [1]. The trees used in the current system cluster on immediate phonetic context and immediate word-boundary context, and have 4767 leaves. The speech generated therefore has ideal phonetic and prosodic targets, but uses tree-based non-contiguous segments and performs prosody modification.

## 2.6. System 6

System 6 speech introduces the use of synthetic duration target values. The speech is generated in exactly the same way as System 5 speech except that the target duration values are obtained from the standard TTS rule-based front-end instead of from the System 1 speech. Pitch and energy target values still come from System 1.

## 2.7. System 7

System 7 speech introduces synthetic energy target values. The speech is generated in exactly the same way as System 6 speech except that the target energy values are obtained from the standard TTS energy prediction decision trees described in [1]. Pitch target values still come from System 1.

## 2.8. System 8

System 8 speech introduces synthetic pitch target values. The speech is generated in exactly the same way as System 7 speech except that the target pitch values are obtained from the standard TTS rule-based front-end. System 8 speech is almost pure TTS, except that the text normalisation, phrase boundary placement and pronunciation modules have been effectively replaced by using values derived from the System 1 speech.

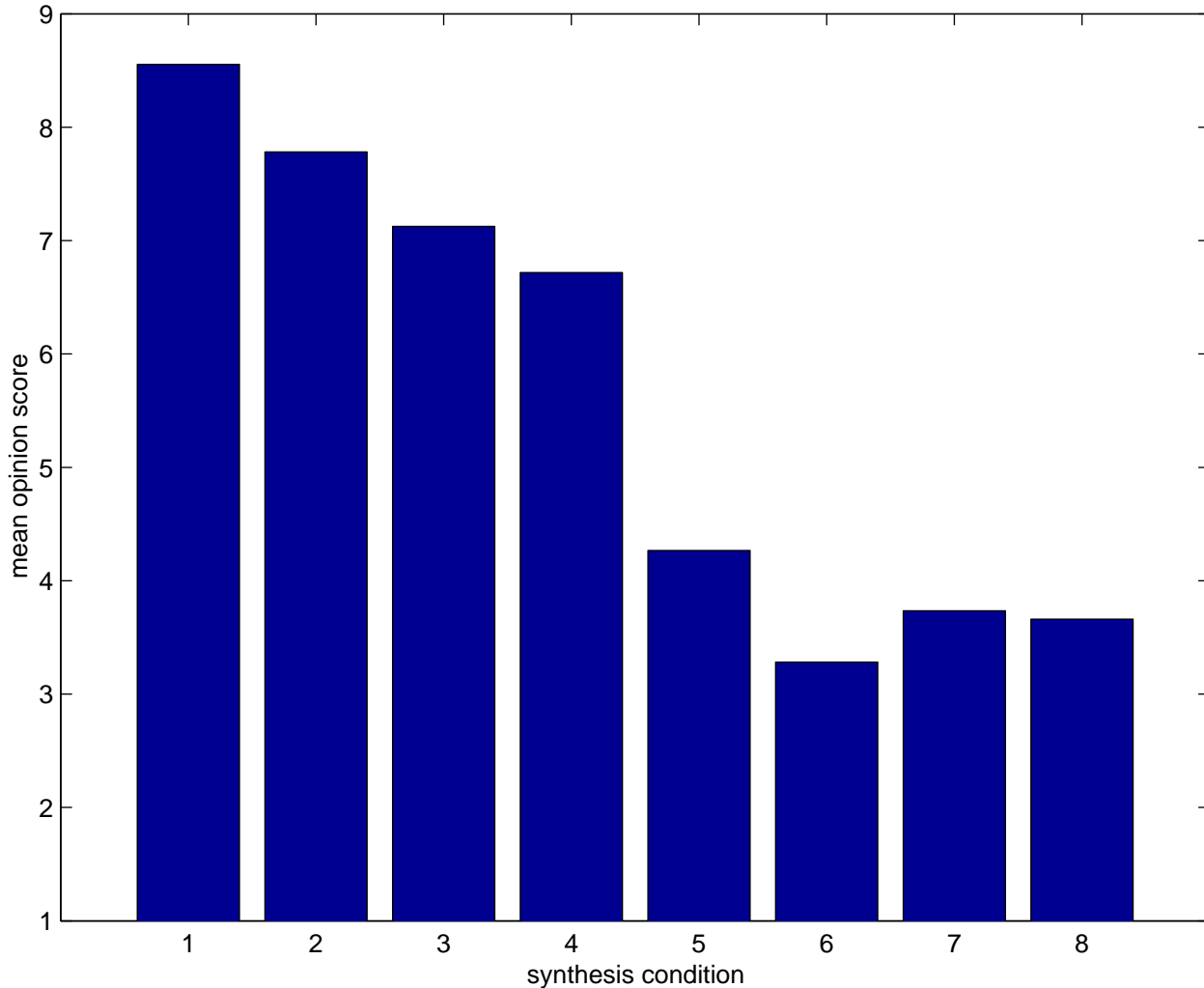


Figure 1: Listening test Mean Opinion Scores for each of the synthesis systems described in the text.

### 3. Listening Tests

The listening test was conducted over loudspeakers in a sound-proofed room. Eight listeners were used, all of whom were native American English speakers, listening to 8 systems synthesising 8 sentences, making for 512 data points in all. The listeners were asked to rate each item they heard on a scale of 1 (bad) to 9 (good). They were asked to judge how good they thought each item was, with no further instructions given on what aspects of the speech to reward or penalise; it was strongly desired that the listeners make up their own minds about what they thought was most important. The listening test presented the eight versions of each sentence as a group, using a different randomised order for each group. The only exception was for the first two items presented, which were from System 1 and System 8 respectively to give the subjects an idea of the quality range they were about to hear. The listeners were instructed to award these items a 9 and a 2 respectively, and the scores for these items were not used to compute the results. The listeners could listen to each item as many times as they liked, and go back and forth between items if they wished.

### 4. Results

Figure 1 shows the Mean Opinion Score (MOS) [2] rating for each system described in Section 2 computed as the mean of 64 data points in the case of Systems 2-7, and 56 data points for Systems 1&8. The data was collapsed across sentences to a grid of system vs listener means, and difference t-tests conducted between the systems. Significant differences were found, at the 5% level using a two-tailed test, between all adjacent systems in Figure 1 except Systems 7&8.

### 5. Discussion

If we ignore for the moment the possibility of inaccuracies in pitch labelling, then we can say that the significant difference between Systems 1&2 is due to degradation associated with the tiny amounts of prosodic modification introduced. We can also say that the significant difference between Systems 2&3 is due to additional prosodic modification by amounts, shown in Table 2, similar in size to those typically seen in synthesis, shown in Table 1. We could therefore deduce that choosing to perform prosodic modification in synthesis, and doing so by the amounts

Number of Context Dimensions Matched	System 5 %age	In leaf %age	Out of Leaf %age
4	28.6	73.4	0
3	47.4	26.2	98.9
2	19.2	0.3	1.1
1	4.3	0.1	0
0	0.5	0	0

Table 3: Percentage of synthesis contexts matching database contexts in different numbers of context dimensions for the segments in (i) the System 5 speech, (ii) the synthesis leaf sequence, and (iii) the complement to the synthesis leaf sequence.

shown in Table 1, is effectively responsible for all the degradation between Systems 1&3. In practice pitch labelling errors may have corrupted the pitch target values extracted from the System 1 speech and the pitch values of the segments used to construct the System 3 speech. However, the laryngograph processing used for pitch determination is extremely robust, and the speaker used was not prone to vocal fry. Only one error in the target pitch values used to synthesise the System 2 speech is audible to the author, and after discounting this sentence Systems 1&2 are still significantly different. It therefore seems likely that most of the degradation seen is due to prosody modification, a result consistent with [3].

The significant difference between Systems 3&4 indicates that even concatenating segments from ideal contexts in very similar readings of the same sentence causes degradation. The degradation may be due to formant discontinuities at the concatenation boundaries, or perhaps due to the concatenation of segments with different original pitch or energy values, or perhaps due to more subtle differences between non-contiguous segments. The result demonstrates that using longer synthesis units is preferable when they are available. However, the relative size of the degradation shown in Figure 1 suggests that this preference should be a relatively low priority.

The largest difference in Figure 1 is between Systems 4&5, on the introduction of the use of decision trees to identify segments in appropriate contexts instead of using segments from ideal contexts. As mentioned in Section 2.5 the decision trees cluster on immediate phonetic context and immediate word-boundary context only. Looking at these four dimensions of context, a simple scoring mechanism is to count the number of context dimensions in which an exact match between the segment's database context and the synthesis context is achieved. Table 3 shows the percentage of states with each score when each of the 1055 synthesis states used in the listening test is compared to (i) the corresponding segment in the speech synthesised by System 5, (ii) the highest scoring of all the segments in the corresponding leaf in the synthesis leaf sequence, and (iii) the highest scoring of all the segments in all the leaves except for the corresponding leaf in the synthesis leaf sequence.

The figures in Table 3 illustrate a number of failings of the decision tree approach to the segment search. Firstly, comparing the 1st and 2nd results columns, it is clear that segments with much better matching contexts were often available from the synthesis leaf than were actually selected from it when synthesising the System 5 speech. Most likely these segments were not selected because other segments had more appropriate prosody or lower spectral continuity costs, [1]. The failing of the decision tree approach is that within a leaf all segments are considered to have an equally good context, which is in fact

not true, and that segments with inferior context matches may then be selected due to potentially small prosodic or continuity advantages. A second failing of the decision tree approach is suggested by the 3rd results column in Table 3 in which it can be seen that a very large fraction of the states to be synthesised had high-scoring segments in leaves other than the one to be used in synthesis. Although the aim in using decision trees is that the leaf descended to will match (perhaps in a broad phonetic class sense) the most important aspects of the synthesis context, this aim is not always achieved [1]. In general, any group of segments with a common context which is not important enough to cause a split during tree building and which is in a different context dimension from that on which the chosen split is based will be fragmented. The result is that highly appropriate segments may often be found in leaves other than the one descended to during synthesis which cannot be used.

Another limitation of the decision trees used in the current study is that they are built from only immediate phonetic and word-boundary contexts. Increased context size has brought improvements to speech recognition systems using decision trees, [4], and may bring improvements in speech synthesis.

The significant difference between Systems 5&6 indicates that synthetic duration targets are inferior to natural duration targets, as might be expected. The significant difference between Systems 6&7 however is surprising, and given that these systems' only difference was their target energies, and that these were very similar (see Section 2.1) it is difficult to explain. Also surprising is the lack of difference between Systems 7&8 on the introduction of synthetic pitch targets, since the speech produced by the two systems was quite noticeably different.

## 6. Conclusions

Three significant conclusions can be drawn from this study:

- The greatest room for improvement is with the segment search. Although decision trees enable the rapid identification of segments to use in a given context, there are a number of shortcomings with the method and better alternatives may exist.
- The prosodic modification typically applied during synthesis, shown in Table 1, significantly degrades the speech quality.
- Minimising the number of concatenation points in synthesis is worthwhile, though should be a relatively low priority.

## 7. Acknowledgments

Thanks to Mike Monkowski for recording the speech data used in this work. Thanks to the members of the IBM speech group who performed the listening tests.

## 8. References

- [1] Donovan, R.E., and Eide, E.M. (1998) The IBM Trainable Speech Synthesis System, *Proc. ICSLP'98, Sydney*.
- [2] IEEE (1969) IEEE Recommended Practice for Speech Quality Measurements, *IEEE Transactions on Audio and Electroacoustics*, Vol. AU-17, No. 3, pp. 225–246.
- [3] Syrdal A.K., Conkie, A. and Stylianou Y. (1998) Exploration of Acoustic Correlates in Speaker Selection for Concatenative Synthesis, *Proc. ICSLP'98, Sydney*.
- [4] Odell, J.J. (1995) The Use of Context in Large Vocabulary Speech Recognition, *PhD. Thesis, Cambridge University Engineering Department*.