RECENT IMPROVEMENTS TO THE IBM TRAINABLE SPEECH SYNTHESIS SYSTEM

E. Eide, A. Aaron, R. Bakis, P. Cohen, R. Donovan, W. Hamza, T. Mathes, M. Picheny, M. Polkosky, M. Smith*, M. Viswanathan*

IBM T.J. Watson Research Center, Yorktown Heights, NY 10598 U.S.A.
* IBM, Boca Raton, FL 33487 U.S.A.
{ eeide,asaaron,bakis,pscohen,red11,hamzaw,nbopro,picheny,polkosky,smithme,maheshv} @us.ibm.com

ABSTRACT

In this paper we describe the current status of the trainable text-to-speech system at IBM. Recent algorithmic and database changes to the system have led to significant gains in the output quality. On the algorithms side, we have introduced statistical models for predicting pitch and duration targets which replace the rule-based target generation previously employed. Additionally, we have changed the cost function and the search strategy, introduced a post-search pitch smoothing algorithm, and improved our method of preselection. Through the combined data and algorithmic contributions, we have been able to significantly improve ($p < 0.0001$) the mean opinion score (MOS) of our female voice, from 3.68 to 4.85 when heard over speakers and to 5.42 when heard over the telephone (seven point scale).

1. INTRODUCTION

The IBM text-to-speech system is composed of three major components: a front-end which does text normalization and pronunciation generation, a prosody module which generates pitch, duration, and energy targets, and a back-end which searches a large database to select segments which minimize a cost function, concatenates them, and optionally performs signal processing on the resulting synthetic speech. All three major components are undergoing continual improvements. In this paper we describe improvements made to the prosody module, to the back-end, to the database which the back-end searches, and to the preselsection method for offline selection of segments from that database to be considered in the search. All of the algorithmic improvements mentioned, as well as the procedures we followed for collecting new databases, will be described.

2. RECENT ALGORITHMIC IMPROVEMENTS IN THE IBM TTS SYSTEM

The previous version of our concatenative synthesis system is described in detail in [1], and summarized briefly here.

During synthesis, text processing, text-to-phone conversion, and phrase boundary placement are performed by an independent rule-based front-end. Previously, this rule-based front-end was also used to generate duration targets and F0 targets; new statistical methods for predicting these prosodic targets have been implemented to replace the rule-based targets. The results of the front-end processing and the prosodic target generation are passed, one phrase at a time, to the back-end, which generates the synthetic speech by selecting units to minimize a cost function. The system uses subphoneme-sized segments as its basic synthesis units, which correspond to states in a hidden Markov model. Decision trees [2] for context definition are used in conjunction with a dynamic programming segment search.

We have developed several algorithms, described below, which enhanced the quality of the synthetic speech produced by the system. We replaced the rule-based system for generating F0 contours with a statistical approach using a decision tree. We also replaced the rule-based system for producing duration targets with a decision tree approach. We changed the cost function used in the segment search as well as the search strategy itself. We introduced an algorithm for smoothing the pitch contour which results from concatenating the segments selected for synthesis. Finally, we improved our algorithm for preselecting segments to be considered in the search.

2.1. F0 Target Estimation

We replaced the rule-based system for generating pitch target contours with a decision tree model. In this methodology, an end pitch and a delta pitch for each syllable are predicted from a set of features gathered from the text. Features include:

- Lexical stress of the current syllable
- Phrase level stress of the current word, as predicted by the rule-based front-end processor
- Distance of the current word from the beginning of the current phrase
- Distance of the current word from the end of the phrase
- Part of speech of the current word

For each syllable, the feature vector associated with that syllable along with the feature vectors associated with the two syllables to the left and to the right are concatenated, and associated with an observation vector consisting of $log(p)$ and Δp, where p is the pitch in Hertz at the end of the syllable nucleus. From these feature vectors and observations, a decision tree is built to maximize the likelihood of the observations.

During synthesis, the same features are assembled and dropped down the tree for each syllable. The mean pitch and mean delta pitch at the resulting leaf are used to construct the target pitch contour. The estimated end pitch and the delta pitch of the syllable are used to calculate a target start pitch and end pitch for each segment, which are used to evaluate the pitch target component of the cost function for each database segment under consideration for selection.

The new pitch targets, when accompanied by a corresponding change to the cost function which is described in section 2.3, lead to a significant improvement in performance as shown in table 1.

2.2. Target Duration Estimation

Analogously to the F0 contour generation, we have replaced the rule-based approach for generating duration targets with a decision tree model. For each phone to be synthesized, a set of features are derived from the text. Features include:

- The phoneme identity, as well as of the two phones to the left and to the right of the current phone
- The voicing (voiced/unvoiced) of the current phone and of the two phones to the left and to the right of the current phone
- The broad class of the current phone (vowel, semi-vowel, fricative, nasal, plosive) as well as of the two phones to the left and to the right of the current phone
- The total number of syllables in the word to which the current phone belongs
- The syllable number of the syllable to which the current phone belongs within the word
- The total number of syllables in the word minus the current syllable number
- The lexical stress of the current syllable
- The phrase-level stress of the current word, as predicted by the rule-based front-end
- The distance of the current word to the beginning of the current phrase
- The distance of the current word to the end of the phrase
- The part of speech of the current word
- The type of the current phrase (yes/no question, "wh" question, comma, period, etc.)

These features are then paired with the observation $log(d)$, where d is the duration of the current phone. From the feature vector and observation pairs, a decision tree is constructed to maximize the likelihood of the observations assuming a Gaussian distribution at each node of the tree.

In synthesis, feature vectors are compiled from the text to be synthesized in the same manner as was used for training the decision tree. Those feature vectors are then dropped down the tree; the mean of the duration of all training vectors mapping to that leaf is then used as the target duration for the phone to be synthesized.

The duration tree was shown to provide a significant improvement over the rule-based durations when combined with pitch smoothing, to be described in section 2.5, as shown in table 2.

2.3. Cost Function

The cost function mentioned in [1] was designed under the assumption that pitch and duration modification would be performed to force the synthetic speech output to reach the prosodic targets. Thus, cost curves were designed through trial-and-error to reflect the amount of audible degradation introduced by modifying the segment's inherent prosody by different amounts. For example, a database segment with a duration longer than the target was not penalized, while a segment with a duration shorter than the target was penalized, because modification to shorten durations introduces fewer artifacts and is generally preferred over performing modification to lengthen durations.

However, as the signal processing necessary to alter the prosodic content of the speech segments introduces undesirable artifacts into the signal that cause the speech to sound unnatural, we decided to minimize the amount of signal processing we would do, and we redesigned the cost function under the assumption that prosodic modification would *not* be performed. Thus, rather than penalizing segments based on how much distortion would be introduced if that segment were modified to match the target, we penalize segments based on the distance between the target and the database segment value. Furthermore, we introduced tunable weights on each component of the cost function so that, through tradeoffs, a desired attribute may be approximated more closely, e.g. better spectral smoothness at the expense of achieving the pitch target. After the database segments which minimize the cost function are selected, we do perform prosodic modification, forcing the speech to follow a piecewise linear pitch contour. These contours linearly join the end pitches of the selected segment pitches, rather than the target pitches. They have the property that they do not exhibit sudden jumps in pitch and at the same time they reduce the amount of distortion introduced to the signal through signal processing relative to using the target pitches, since they typically require only small modifications to the original pitch.

We performed a formal listening test (seven-point scale), asking participants to rate the naturalness of speech generated from the rule-based pitch target generation and the accompanying cost function (pitch modification performed), and the decision-tree pitch target generation and the new cost function (modification of pitch to the target not performed). The results are shown in table 1. The improvement from system A to system B is statistically significant ($p = 0.002$).

System	MOS
A. Rule-based pitch, old cost function	2.6
B. Decision-tree pitch, new cost function	4.6

Table 1. *Results of a listening test considering rule-based vs. decision-tree based pitch targets.*

The piecewise linear pitch contour does suffer from discontinuities in the derivative of the pitch, when the linear segments join at sharp angles. Rapid changes in the slopes of linear segments can lead to a "warble" heard in the synthetic speech which is quite unpleasant. To overcome this problem, a further enhancement of the piecewise linear pitch contour is achieved through the pitch smoothing algorithm described in section 2.5.

2.4. Search

The search for the sequence of speech segments which will minimize the cost function is at the heart of a unit-selection based speech synthesis system and consumes the bulk of the processing time. As the amount of data grows, the search time can grow exponentially. Thus, the need for fast pruning algorithms which preserve the minimum-cost solution presents itself. The previous version of our system did pruning of candidates based on their proximity to the prosodic targets; the nearest N (typically 5) candidates to the targets were retained, and a full dynamic programming search of the resulting grid was performed. However, we realized that numerous acceptable prosodic renditions of a given text may be possible and that we may want to relax the prosodic constraints while enforcing spectral smoothness. Thus we changed the pruned-grid search to a Viterbi beam search [3], in which all candidates in the target leaf, as defined by the decision tree, are considered. For speed, we no longer consider segments from other contexts, with some penalty, as was described in [1]. All segments in the target leaf are scored using the best cost among the costs computed from all possible predecessors' costs to date and

the pitch transition and spectral transition costs to them. All segments within twice the beamwidth of the best segment are then scored against the target prosody. Segments within the beamwidth of the best segment are then retained for use in the next step in the dynamic programming forward pass.

2.5. Pitch Smoothing

After segment selection takes place, a pitch contour consisting of a piecewise linear connection of the observed end pitch of each selected segment is constructed. By using the observed pitch rather than the target pitch, we avoid making large modifications to the original signal, thereby introducing less distortion to the synthetic speech. Although this method of constructing a pitch contour from concatenated segments is continuous in the mathematical sense, it can exhibit rapid fluctuations and sharp corners (large discontinuities in the first derivative) which sound as if the talker were shaking or distressed. We introduce additional smoothing to remove this effect.

Smoothing is accomplished by convolving the piecewise linear pitch contour with a kernel function. In order to avoid a displacement of the contour in time, the kernel is made symmetric, so that the phase shift is zero. Good results were obtained with the double exponential kernel $h(\tau) = \frac{1}{2\tau_0} e^{-\frac{|\tau|}{\tau_0}}$, with values of τ_0 in the range of 30-90 msec.

Combined use of the pitch smoothing and decision-tree-based duration models were shown to significantly outperform the baseline system in a listening test (five-point scale), as shown in table 2. The improvement from system A to system C is statistically significant ($p < 0.05$).

System	MOS
A. Piecewise linear pitch, rule-based durations	3.42
B. Smoothed pitch, rule-based durations	3.50
C. Smoothed pitch , decision-tree durations	3.63

Table 2. *Results of a listening test considering pitch smoothing and decision tree based duration modeling.*

2.6. Preselection

A further improvement to the IBM TTS system came in the method which we use for preselection. The previous version of the system retained only the first N (typically 25) segments for each leaf. A new, data-driven algorithm was implemented which retained segments based on the number of times they were used in synthesizing a large test corpus. See [4] for details.

3. RECENT IMPROVEMENTS TO THE DATA

In addition to the algorithmic changes described above, we have also recorded a new female voice. The decision to undertake this data collection was driven by feedback from our customers, whose comments on the previous voices indicated that the dynamic range of the pitch was too small, leading to dull, disinterested sounding voices. We attempted to increase the dynamic range of the pitch by scaling the predicted pitch targets, and either trying to find segments that had similar pitch and do no signal processing, or by adjusting the pitch of the original segment to match the desired target. Both of these methods failed to improve quality, however. When we tried to find segments whose pitch matched the adjusted targets, we could not, because the underlying database consisted of segments with very little pitch variation. Thus, when *not* modifying the segment pitch to match the target pitch, there was little perceived difference whether or not the dynamic range of the targets was scaled, since there were no segments with the pitch of the scaled target to be found. Also, when we tried doing signal processing to force the selected segment pitch to match the target, we introduced many unnatural artifacts since the targets and the segment pitches were often very different.

Thus, we came to the conclusion that we needed to record new, "enthusiastic" voices, having previously established that recording a given style of natural speech leads to synthetic speech exhibiting that same speaking style [5].

3.1. Auditions

In order to select the speakers for our new voices, we held a series of auditions. The idea of auditioning speakers for TTS was presented in [6]. Our first audition consisted of recording about two minutes of speech from each of 25 female speakers. We then created a telephone demonstration showcasing natural speech from each of those voices, plus our original system's speaker, and asked a large number of listeners to call in and rate the voices. From those votes, we selected six finalists for an "extended audition" which consisted of recording 1400 phonetically-balanced sentences. From those recordings, we built text-to-speech systems for each of the six finalists, and again asked listeners to call in and rate the synthetic voices. From that vote we selected the top vote-getter as our new female voice.

3.2. Scripts

In addition to a new speaker, we adopted a new recording script, which was much larger than the previous one. The old script consisted of 1400 phonetically-balanced sentences, which were constructed for the purpose of obtaining a compact database. However, those data lacked many phrases which were likely to be synthesized in real-world applications, such as "Welcome to ..." Thus we constructed a new script; the first 1400 were the same phonetically balanced sentences used originally and the remainder were taken from a variety of different domains, including news reports from wire services, e-mail, airline phone prompts, finance-related material, dates, weather, and numbers.

The total script corresponds to about 15 hours of speech, including silence, or nearly 11 hours of speech excluding silence. This is roughly ten times larger than our previous database.

3.3. Recordings

Recordings were made in a professional recording studio in which the speaker and recordists were in separate rooms connected by a window. Text-to-speech recordings require, in addition to a low noise floor, extraordinarily low reflections in the booth; thus the surfaces of the recording room (excluding the window) were covered in absorbent material and the speaker was speaking parallel to the window so as to minimize reflections. Recording sessions typically lasted about four hours per day, with several breaks to avoid speaker fatigue. Direction was provided to the speakers to elicit a warm and friendly speaking style. Measures were taken to ensure consistency in volume and style within and between sessions.

4. OVERALL RESULTS

Through the combination of improved algorithms, a change of speaker and speaking style, and an increase in the amount of recorded data, we have been able to significantly increase the intelligibility, naturalness, prosody, and social impression [7] of our female voice when it is heard on desktop speakers (as in web-based applications) or on the telephone (as in telephony-based applications). We have improved the overall MOS-X score from 3.68 for the previous system to 4.85 for the new system via speakers and 5.42 via telephone, a significant increase in listener perception of the new voice in both listening environments ($p < 0.0001$).

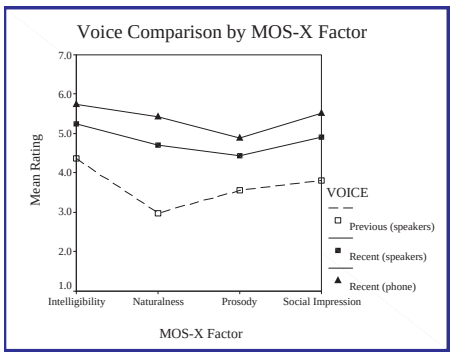

Fig. 1. Results of a listening test (7-point scale) comparing the previous and most recent female voice in the IBM U.S. English text-to-speech system, heard via desktop speakers and telephone. Improvement from Previous to Recent voice is statistically significant across all MOS-X factors ($p < 0.0001$).

Interestingly, when we compared the new female speaker to our previous female speaker, using the same amount of data (1400 phonetically balanced sentences), we actually observed a large drop in performance, from an MOS score of 3.8 for the previous voice to a score of 3.0 for the new voice, which was statistically significant ($p < 0.05$). This result may be explained by the fact that the new voice has a much larger prosodic range and therefore requires much more data to adequately cover the space. With the small amount of data used to build the systems in this test, large spectral and pitch discontinuities were observed in the synthetic speech; the discontinuities diminish as the dataset size increases.

In addition to U.S. English, the algorithmic and data collection procedures have been tested in other languages. The algorithms result in statistically significant improvements in the intelligibility, naturalness, prosody, and social impression of French ($p < 0.008$, all MOS-X factors) and German ($p < 0.015$, all MOS-X factors), as compared with the previous systems. The overall MOS-X score also revealed substantial perceptual improvement for French ($p < 0.001$) and German ($p < 0.0001$).

Finally, we note that we have been able to achieve a reduction in the average number of non-contiguous segment concatenations, from a previous value of about 70% (that is, out of every ten possible places for a non-contiguous concatenation to occur, seven did) to a current value of about 30%. This reduction was effected primarily through the changes to the cost function and search strategy, and through the use of additional data.

5. ACKNOWLEDGEMENTS

The authors would like to acknowledge the work of Bernhard Zeller of IBM Germany and Stephane Revelin of IBM France, who have led efforts to build corresponding text-to-speech synthesis systems in their respective languages, and we thank them for sharing their listening test results. Thanks to Doris Komarek of IBM Austria for running the listening tests in those languages.

Special thanks to Jaime Botella Ordinas of IBM Spain for his invaluable assistance in the development of the TTS code. Thanks also to Jorge Gonzalez Lopez, Volker Fischer, Marion Mast, Julia Vogl, Darren Green, Joachim Premm, Jochen Friedrich, James Lewis, Allen Delmar, Philip Gleason, and Tom Rutherfoord for their many contributions to the IBM synthesis project. Finally, thanks to the studio and professional speakers involved in the recording process, and to all the listeners who participated in the many evaluations.

6. REFERENCES

[1] Donovan, R.E., et al. “Current Status of the IBM Trainable Speech Synthesis System,” Proc. 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Atholl Palace Hotel, Scotland, 2001. Available from http://www.ssw4.org

[2] Breiman, L., et al. “Classification and Regression Trees.” Chapman & Hall. 1984.

[3] Forney, G.D. “The Viterbi Algorithm,” Proc. IEEE, vol 61, pp 268-278. 1973.

[4] Hamza, W. and R. Donovan. “Data-Driven Segment Preselection in the IBM Trainable Speech Synthesis System.” Proc. ICSLP. Denver, CO. 2002.

[5] Eide, E. “Preservation, Identification, and Use of Emotion in a Text-to-speech System.” IEEE Workshop on Speech Synthesis. Santa Monica, CA. September, 2002.

[6] Syrdal, A., A. Conkie, and Y. Stylianou. “Exploration of Acoustic Correlates in Speaker Selection for Concatenative Synthesis.” Proc. ICSLP, Sydney,Australia, 1998.

[7] Polkosky, M. and J. Lewis. (in press). “Expanding the MOS: Development and psychometric evaluation of the MOS-R and MOS-X.” International Journal of Speech Technology.