ON THE RELATIVE IMPORTANCE OF DIFFERENT PROSODIC FACTORS FOR IMPROVING SPEECH SYNTHESIS

I.Bulyko[†], M.Ostendorf[†], and P.Price[‡]
[†]*Boston University, Boston, MA*
[‡]*SRI International, Menlo Park, CA*

ABSTRACT

We present results of perceptual experiments geared toward assessing the relative importance of several prosodic factors in synthetic speech, showing that naturalness, relative to a target speaking style, can be significantly improved through both symbolic label prediction and better F0 and duration generation. Our experiments utilized a novel perceptual experiment paradigm, where we supply each test subject with two reference utterances in order to obtain reliable absolute scores that indicate magnitude of improvement. The approach gives ratings that are comparable across experiments. Results also show a strong interaction between detailed F0 and duration controls.

1. INTRODUCTION

A growing number of speech recognition applications are creating an increasing demand for better quality speech output. Further, the possibility of generating speech from "concept" provides the opportunity for prosody to play an even more important role, not only in improving naturalness and intelligibility, but also in contributing to the perception of a particular speaking style. Many aspects of prosody could be improved, from placement of symbolic accent and phrase boundary markers to control of continuously varying parameters such as phone duration and fundamental frequency. In this paper we will assess relative contributions of different aspects of prosody towards improvements in achieving a target speaking style in the context of concatenative synthesis.

Due to the subjective nature of speech perception, evaluation of synthetic speech is a difficult problem [8], and has been the subject of ongoing debate. A widely used approach to measuring speech naturalness is to ask the subjects for an indication whether one utterance was better, equal, or worse than another [2, 4]. Even though this relative scale reliably indicates differences, it fails to show magnitude of improvement. In this work we implement a perceptual experiment paradigm that involves supplying each test subject with two reference utterances, thereby making the scoring scale more quantitative and comparable across experiments.

2. PERCEPTUAL EXPERIMENTS

2.1. General Method

Experiments were conducted in which, twelve naïve subjects, all native speakers of American English, listened to several versions of each of eight synthetic utterances and scored them on a 1-10 scale. The target speech was from a corpus of radio news stories, and all of the utterances were generated by the Entropic TrueTalk speech synthesizer [9] in a 16-bit 16kHz format.

For each sentence, listeners were provided with two reference versions of the sentence and four versions to score. The references included the synthesizer's default text-to-speech (score 1 = least natural) and a version synthesized with natural phone durations and F0 contour as measured from a version of the sentence spoken in an actual radio news broadcast (score 10 = most natural). The use of reference versions was intended to help the subjects focus attention on prosodic differences rather than segmental quality, which would be the same in all six versions. Any pronunciation errors made by the synthesizer were corrected in all versions, again to focus on prosodic factors. The use of references was also intended to reduce the subject variability in scoring and to suggest to the listeners the target speaking style as a more concrete definition of “natural”.

Since it was too tedious for subjects to rate several versions of an utterance at once, two experiments were run. There was some overlap in the stimuli to test whether the scores would be similar across experiments. Four subjects participated in both experiments, but the experiments were separated in time by a few months.

Subjects were asked to listen to the reference utterances first and assume those utterance were given scores of 1 (least natural) and 10 (most natural). Then they could listen to the test utterances and score each for naturalness on a scale of 1 to 10. The four versions of each test utterance were arranged in random order to account for learning bias, however the order of presentation of the eight sentences preserved the flow of the discourse in the original news story so that the target prosodic style would indeed seem appropriate or “natural”. Subjects were allowed to play the test and reference utterances as many times as they wanted. Listening was performed via loudspeakers in an isolated room.

2.2 Prosodic Control Variables

The prosodic parameters that we allowed to vary in our experiments include symbolic labels (phrase breaks, pitch accents and tones) and acoustic parameters (phone duration, pitch range and F0 contour).

For the stimuli where “natural” symbolic labels were used, these were based on a hand-labeled prosodic transcription of the target utterances based on the ToBI labeling system [5]. Phrase breaks included location of minor and major phrases (ToBI breaks levels 3 and 4, respectively). For the cases where breaks and accents are used, but no tones, the synthesizer default tone assignment is implemented. For the case where the ToBI tones

are used, the bitonal accents (L+H* and L+!H*) are converted to single tones (H* and !H*) to be consistent with the F0 prediction algorithm, described shortly. When the tones are used in the natural or generated F0 contour, but not explicitly input to the synthesizer, we refer to them as "implicit".

The natural phone durations were based on a phone segmentation derived from Viterbi alignment of the target utterances using a speech recognition system and then hand corrected.

Natural pitch range and F0 contour controls were based on F0 values measured using the Entropic Waves pitch tracker with an F0 sample rate of 100Hz. The F0 contour was then smoothed using a 5 point median filter. The pitch range was estimated for each minor prosodic phrase as the maximum value within the phrase, using hand-labeled phrase boundaries. The "natural" F0 contour was hand-corrected in places where F0 tracking errors caused a noticeable difference in the prosody of the spoken and synthesized versions. Natural pitch range is implicit in the natural F0 contour, as are tone labels and segment durations.

The synthesized versions using the predicted F0 contour are based on an automatically trained model that separately represents range, tone and segmental effects [7]. In all cases, the range is "natural" and the hand-labeled tones are input to the f0 prediction algorithm. The algorithm also requires phone durations, which can be natural or the synthesizer defaults for the prosodic context.

2.3. Specific Experiments

Type	Phone Duration	Breaks, Accents	Tones	Pitch Range	F0 Contour
Series A	Default	Natural	Default	Default	Default
Series B	Natural	Natural	Default	Default	Default
Series C	Default	Natural	Default	Natural	Default
Series D	Natural	Natural	Natural	Natural	Predicted

Table 1. Prosodic controls for different versions in Experiment 1.

Type	Phone Duration	Breaks, Accents	Tones	Pitch Range	F0 Contour
Series A	Default	Natural	Default	Default	Default
Series B	Default	Natural	Natural	Natural	Predicted
Series C	Natural	Natural	Natural	Default	Default
Series D	Natural	Natural	Natural	Natural	Predicted

Table 2. Prosodic controls for different versions in Experiment 2.

Two experiments were conducted with different prosodic controls manipulated in each case, as summarized in Tables 1 and 2. The specific variations chosen aimed at assessing the potential for improved perceived naturalness, as well as separating out the role of the different factors that together seemed to give good performance of the automatic F0 prediction algorithm. The reference utterances were the same in both cases, using the default synthesizer output (with pronunciation corrections) as reference 1 and the version with natural durations and F0 contour as reference 10. The breaks, accent/tones, and pitch range controls for reference 10 are implicit in supplying natural durations and the F0 contour. The series A and D utterances are similar in both experiments to test for consistency across experiments.

3. RESULTS

The results of the two experiments are shown in Figures 1-4. Figures 1 and 2 indicate the average ratings of the different versions, and Figures 3 and 4 give the distribution of scores accumulated over all utterances and listeners. With the exception of version B in Experiment 1, which has an atypical bimodal distribution that will be discussed later, all cases had a single peak in the distribution and standard deviations in the range of 1.7-2.3. The standard deviations were lower for the second experiment, suggesting that either the subjects were more consistent and/or the task was easier.

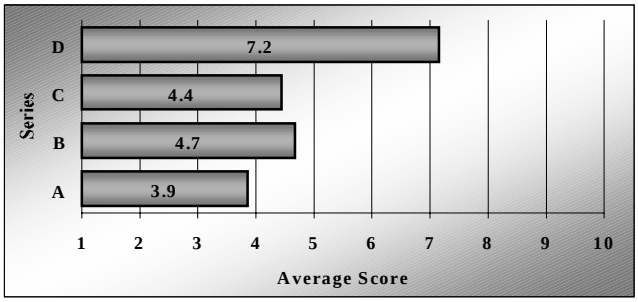


Figure 1. Average scores – Experiment 1.

Using a paired difference t-test for assessing significance of response differences within an experiment, we found that the differences between all versions with respect to each other and the references are significant with $p<10^{-4}$, with the following exceptions. The differences between versions B and C in both experiments are not significant, and the difference between A and C in the first test is only marginally significant with $p<.05$.

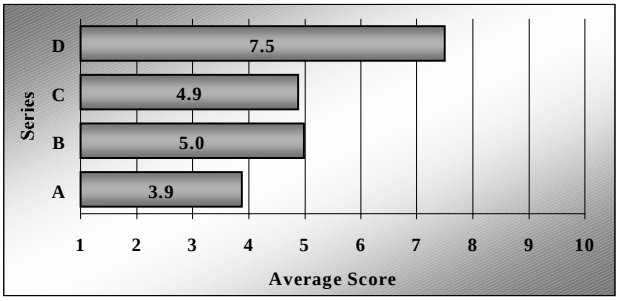


Figure 2. Average scores – Experiment 2.

		Expt 2			
		A	B	C	D
Expt 1	A	n.s.	0.1	0.01	0.0001
	B	n.s.	n.s.	n.s.	0.001
	C	n.s.	n.s.	n.s.	0.001
	D	0.0001	0.001	0.001	n.s

Table 3. Significance of cross experiment differences.

Table 3 gives the statistical significance of differences

between the responses for different versions, comparing across tests. We find that there is no significant difference between the responses to the two version A sets, as one would hope since these utterances are identical. Similarly, the differences between the two version D sets are not significant, since these utterances are essentially the same[1].

It is possible to translate the results into relative rankings by counting the number of times one version is rated higher, lower or tied with another version, as illustrated in Figure 5 for Experiment 1. Using relative rankings leads to similar conclusions about which differences are significant within an experiment, but would not allow comparisons across experiments.

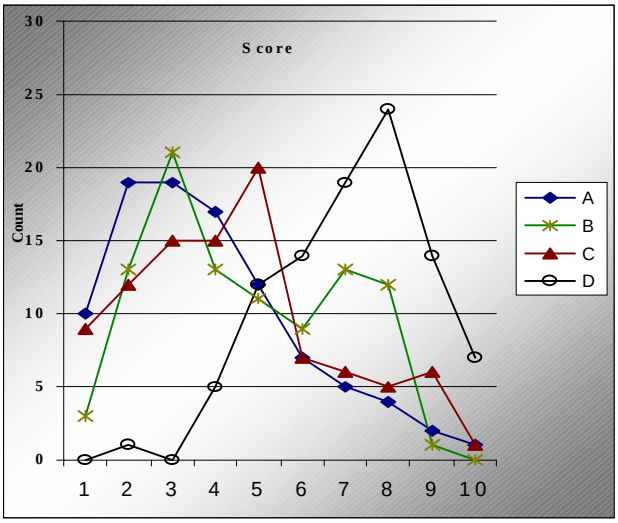


Figure 3. Distribution of scores – Experiment 1.

4. DISCUSSION

The first conclusion that we draw from these experiments is that, using the method of comparing to low and high references, the absolute scale rating can be effective, making it possible to compare results across experiments. As evidence, we take the similarity of the ratings of versions A and D in both experiments. This result is important, since previous work with other paradigms has shown that ratings do not generalize. There has been criticism [2] of absolute scoring scale for not being reliable in a framework of a perceptual experiment. However, this criticism is based on a paradigm that does not use two references, and our results demonstrate that the absolute scale can be reliable, in which case it is more powerful than relative rankings.

Second, we find that improved prediction of symbolic labels can lead to higher perceived naturalness ratings, at least in terms of selecting the appropriate location based on the significance of the difference between version A and reference 1 in both experiments. This is in contrast to the results reported in [6] on similar data, because of the difference in experiment design. In particular, the fact that this experiment focused on achieving a particular style rather than "naturalness" defined in a more abstract sense, is probably the main factor. The contribution of tones (versions B1 vs. C2) is a much smaller effect than that of placement, which is not significant in our experiments.

Approximately equal scores for B2 and C2 (5.0/4.9) (Figure 2) suggest that natural phone durations and predicted F0 contour each make an equal contribution to the naturalness of the utterances.

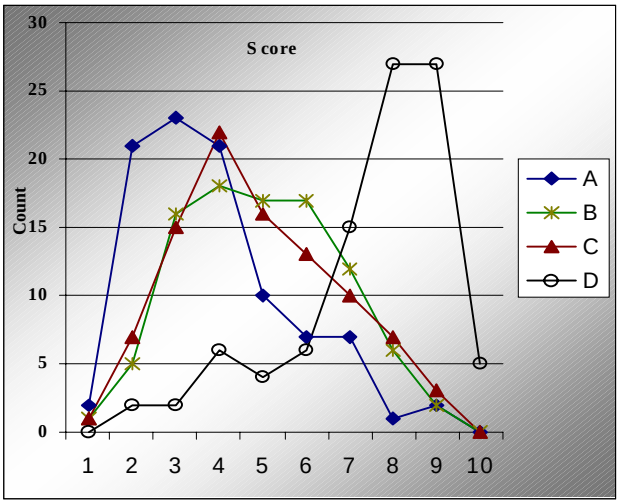
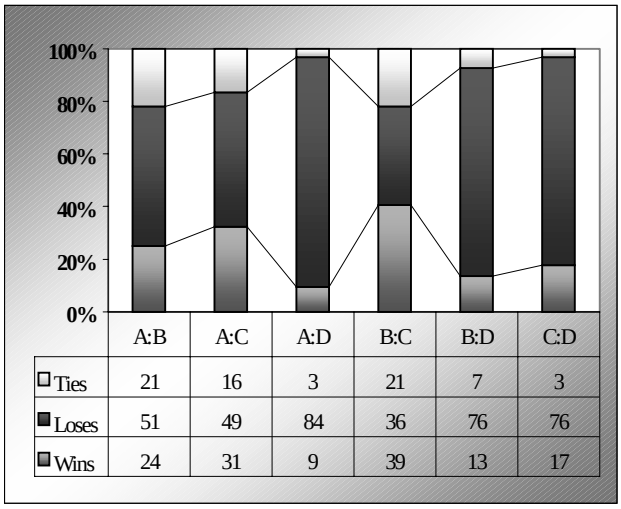


Figure 4. Distribution of scores – Experiment 2.

The automatic F0 prediction algorithm was given a high rating (7.2 and 7.5 in the series D1 and D2 experiments, respectively), as would be expected from the results reported in [7]. However, these utterances used natural durations, tone labels and pitch range, which raised the question of how much these other factors contributed to the high rating. From the A1/C1 comparison, which is only marginally significant, we can see that pitch range alone does not explain the high score, though it may be important to have in combination with a detailed F0 model. From the B1/C2 comparison, we see that the use of hand-labeled tones explains only a small amount of the difference. The B1/D1 comparison shows that duration alone does not lead to improved performance (gains of 0.8 vs. 3.3 relative to A1, respectively). However, the A2/B2/D2 comparison shows that the F0 predictor does not get all the credit, since the gain of the F0 model with default durations is much lower than the gain using natural durations (1.1 vs 3.5 difference relative to A2, respectively). Together, these results suggest a strong interaction between F0 and duration, showing gains that are more than additive when the two improvements are combined.

The bimodal distribution of scores for series B1 made us expect a subject split. Indeed, it appeared that subjects could be divided into three categories: those who consistently preferred natural durations over F0 range; those who found F0 range more salient; and those who did not indicate a clear preference. Those who had a preference between versions B1 and C1, gave the same score (5.1) to the version they preferred, and gave the other version a substantially lower score. These results suggest that subjects in these groups may be less sensitive to one

prosodic parameter (duration or F0 range). However, it may also be that the two groups are differently sensitive to artifacts associated with signal modifications. The neutral group (with no definite preference between B1 and C1) gave similar scores to both (4.5/4.3) that were substantially higher than the score for A1 and A2 (3.2).



	A:B	A:C	A:D	B:C	B:D	C:D
Ties	21	16	3	21	7	3
Loses	51	49	84	36	76	76
Wins	24	31	9	39	13	17

Figure 5. Pair-wise score relationships – Experiment 1.

5. CONCLUSIONS

Although we cannot point to one acoustic cue being more important than another, we have shown that perceived quality can be significantly improved through both symbolic label prediction and acoustic parameter generation. In summary, the following major conclusions can be drawn from the results of our perceptual experiments:

- F0 contour prediction has played a significant role in improving the naturalness.
- Symbolic phrase and accent location markers gave significant improvement over the baseline reference, but the additional gain from specific tonal markers was not significant.
- Predicted F0 and natural phone durations separately produced a similar level of perceived improvement. However, when combined, these two prosodic features appeared to amplify each other's contribution.
- Subjects can be categorized by their sensitivity to pitch range and/or phone durations.
- An absolute scoring scale has proved itself to be reliable, when supplied with high and low references.

NOTES

1. In fact, there might be minor differences if the synthesizer uses tone labels to control factors other than F0 and duration, such as energy.

REFERENCES

[1] Auberg, V., Gr pillat, T. and Rilliard, A. 1997. Can we perceive attitudes before the end of the sentences? The gating paradigm for prosodic contours. *Proceedings EuroSpeech*. Rhodes, Greece. 871-874.

[2] Black, A.W. and Taylor, P. 1997. Automatically clustering similar units for unit selection on speech synthesis. *Proceedings EuroSpeech*. Rhodes, Greece. 601-604.

[3] Campbell, N., Itoh, Y.,Ding, W. and Higuchi, N. 1997. Factors affecting perceived quality and intelligibility in the CHATR concatenative speech synthesizer. *Proceedings EuroSpeech.* Rhodes, Greece.

[4] Hogberg, J. 1997, Data driven formant synthesis. *Proceedings EuroSpeech*. Rhodes, Greece. 565-568.

[5] Pitrelli, J., Beckman, M. and Hirschberg, J. 1994. Evaluation of prosodic transcription labeling reliability in the ToBI framework. *Proceedings ICSLP.* 1, 123-126.

[6] Ross, K. and Ostendorf, M. 1996. Prediction of abstract labels for speech synthesis. *Computer Speech and Language.* 10(3), 155-185.

[7] Ross, K. and Ostendorf, M. 1999. A dynamical system model for generating fundamental frequency for speech synthesis. *IEEE Transactions on Speech and Audio Processing*. In press.

[8] Van Santen, J. 1993. Perceptual experiments for diagnostic testing of text-to-speech systems. *Computer Speech and Language*. 7, 49-100.

[9] Entropic Text-to-speech synthesis software, version 2.0. Retrieved March 5, 1999 from the World Wide Web : http://www.entropic.com