

Effects on TTS quality of methods of realizing natural prosodic variations

Matthias Jilka, Ann K. Syrdal, Alistair D. Conkie and David A. Kapilow

AT&T LABS RESEARCH

Florham Park, NJ

E-mail: {jilka, syrdal, adc, dak}@research.att.com

ABSTRACT

This study attempts to determine whether natural prosody variations and different methods of applying prosodic patterns are relevant to listeners' perceptions of synthetic speech quality. The prosodic patterns of five test sentences including Yes-No questions, Wh-questions, declaratives, and continuation rises as produced by six female native speakers of four varieties of English were imposed on the same US English voice using four different methods. Results of a perceptual experiment involving 32 listeners show that the methods resulting in fewer distortions and artifacts are preferred to a significant degree, thus favoring synthesis approaches with minimal signal modification and prosodic patterns without extreme parameter values. An additional test that includes a more obvious prosodic phrasing error further clarifies that prosody becomes a more significant factor when no meaningful interpretation is evident in the given context.

1. INTRODUCTION

As segmental quality in unit selection based concatenative speech synthesis has improved, conspicuous errors involving concatenation discontinuities or bad units are less likely to dominate a listener's perception. As a consequence prosodic features such as intonation and phrasing have gained more significance with respect to their contribution to the overall naturalness and intelligibility of such text-to-speech (TTS) systems. However, it is neither clear how sensitive listeners actually are in their perception of an acceptable intonation contour, nor how effectively various methods realize prosodic patterns. An earlier study [1] reported significantly higher synthetic speech quality when units from a large speech inventory were selected that best matched the prosody predicted by the TTS system and simply concatenated, than when the f_0 and duration of these units were modified with either TD-PSOLA or HNM techniques to match exactly the predicted prosody. However, the study could not determine whether the degradation in speech quality resulted from poor TTS prosody prediction, or from signal modification itself. Another study [2] found significant negative correlations between synthetic speech quality ratings and the extent of f_0 and duration modifications required to realize speakers' natural prosody, but there were other significant acoustic correlates of ratings as well. The current study addresses these issues by using natural prosodic variations and a high

quality Residual Excited Linear Prediction (REL) pitch and time-scaling algorithm that is a pitch synchronous overlap add method applied to the residual-domain of an LPC filter similar to PSOLA-REL [3]. This study explores natural prosodic variations in specific discourse situations, different methods of realizing these variations in synthetic speech, and subjective judgments of the resulting speech quality by native and non-native listeners. A linguistic description of the discourse situations and observed intonation patterns is given. The analysis of the listening test results concentrates on the perception of prosody and its interaction with overall quality as it is influenced by different synthesis methods.

2. METHOD

The following five test sentences were chosen for the perception experiment due their general variety of intonation patterns, as well as the occurrence of specific, particularly distinctive tonal phenomena. They are representative of the four major discourse situations declarative (decl), continuation rise (crise), wh-question (whq) and yes/no-question (ynq).

s1: Oak is the type of wood Dennis likes best.(decl)

s2: It snowed, rained, and hailed the same morning. (crise)

s3: How may I help you? (whq)

s4: How would you like to bill this call? Collect? (whq + ynq)

s5: Do you want to make a collect call? (ynq)

2.1. Tonal descriptions

The natural speech productions of the six speakers provided six natural prosodic variants for each test sentence. Three speakers were native speakers of US English (Voices US1, US2 and US3). UK English (UK1), Australian English (AU1), and Indian-accented English (IE1) were represented by one speaker each. This variety of linguistic backgrounds allows for the occurrence of both very similar f_0 patterns and distinctly different ones. In several cases a deviating prosodic pattern can even consist in a final tone entirely different from the contour generally assumed to be typical for a specific discourse situation (as described e.g. in [4]). Figure 1 shows three f_0 contours of renditions of the wh-question "How may I help you?". While two speakers produce essentially the same intonation pattern (in terms of a ToBI tone label description) with a characteristic final fall,

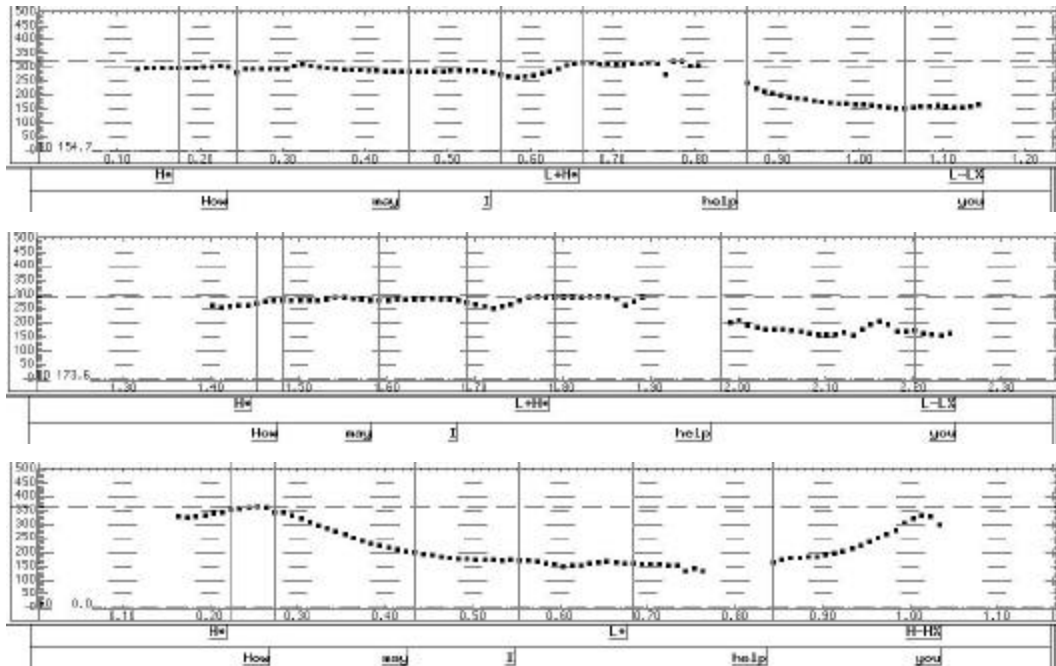


Figure 1. F_0 contours of the utterance “How may I help you” as produced by three different speakers. The top two contours exhibit an identical tune in terms of a ToBI tone label phonological description (final fall in a wh-question), whereas the third one deviates with a strong final rise usually associated with a yes/no-question.

a third one actually realizes the wh-question with a final rise usually reserved for a yes/no-question.

2.2. Methods of prosody realization

The intonation contours and segment durations of these prosodic variants were used to determine the prosody of the test utterances synthesized using each of four different methods.

- **Natural+Warp (N+W):** The continuous utterance spoken by the US English speaker US1 was prosodically warped to match each of the six variants using PSOLA-REL P. In PSOLA-REL P a pitch synchronous LPC analysis is performed on the speech file. The LPC analysis method is derived from the TPC wideband speech coder [5]. The speech is inverse filtered to obtain the residual. For synthesis, time- and pitch-scaling is performed on the residual, and the new synthesized residual is filtered through the time-varying LPC filters. Note that no prosodic modification was required to match this speaker’s utterance with her original prosody, so that one of the six N+W variants of each sentence served as a natural speech control. The N+W technique produces the different prosody patterns without introducing distortions due to concatenation or inappropriate units. Any N+W distortions could only result from PSOLA-REL P processing.
- **Default TTS+Warp (DT+W):** The default synthetic version (using TTS prosody prediction) of each sentence, synthesized with the TTS voice US1, was prosodically warped using PSOLA-REL P to match

each of the six variants. DT+W distortions could result from both concatenation and signal processing.

- **Matched TTS (MT):** A version of each sentence was synthesized using a “unit selection only” technique in which half-phone units were selected by TTS unit selection procedures to best match each of the six prosodic (and phonetic) variants. This was achieved by using as input to unit selection the measured f_0 and duration values, and also the phone parameters (with suitable mappings), from the natural sentences instead of the normal TTS generated parameters. Some f_0 interpolation was performed in order to achieve appropriate f_0 parameters for the unvoiced regions. No prosodic warping of these utterances was performed, so concatenation distortions were possible, but no distortions due to signal processing. This method also differed from both N+W and DT+W methods in that the phones that were specified for unit selection matched the respective original speaker’s phonetic realization and not those produced in that context by the speaker US1 or the US1 TTS output.
- **Matched TTS+Warp (MT+W):** The sentences synthesized using this method combined the latter two approaches, as the MT units were prosodically warped with PSOLA-REL P to precisely match the prosody of the six variants. This approach, like DT+W, is subject to both concatenation and signal modification distortions, but should require less signal modification than DT+W because its units more closely match the desired prosody.

3. TEST DESIGN

Sixteen native speakers of English and 16 non-native speakers of English participated as volunteer listeners in the experiment. All were adults and AT&T employees who were blind as to the identity of the stimuli or the specific purpose of the experiment.

An interactive web-based subjective listening test was conducted in which individual listeners rated each of the 120 test utterances. Ratings were made on a 5-point scale (5=excellent, 4=good, 3=fair, 2=poor, 1=bad). Each listener was presented a different randomized order of test sentences. There were five practice examples preceding the test in order to familiarize listeners with a range of sentences, prosodic variants, and synthesis methods before they began the test. The test typically took approximately 20 minutes to complete.

4. RESULTS AND ANALYSIS

A Group (2) by Method (4) by Sentence (5) by Voice of prosodic pattern (6) repeated measures analysis of variance (ANOVA) was performed, and tests of within-subjects effects (Condition, Sentence, Voice) and between-subjects effects (listener Group) were made for significant main effects or interactions as determined by the ANOVA. There were significant within-subjects main effects for Method ($F(3,90)=64.463$, $p<0.001$), Sentence ($F(4,120)=62.466$, $p<0.001$), and Voice ($F(5,150)=81.912$, $p<0.001$). The main effect for Method reflected the fact that N+W ratings were significantly higher than ratings of the other three methods, and the MT method was rated significantly higher than either DT+W or MT+W, whose ratings didn't differ significantly from each other. The main effect for Sentence resulted from the superior rating of s3 over all other sentences, followed by s2 and s4, which did not differ significantly from each other. s4 and s5 were not significantly different from each other, and s1 was rated significantly lower than the other four sentences overall. Distortion due to prosody warping when pitch range discrepancies between the original speakers' patterns and individual words or longer stretches of speech containing considerably too low or high a pitch in the speech to be modified probably accounted for s1 receiving the lowest quality ratings. The Voice main effect was due to significantly different ratings among the six voices: US1's prosodic pattern was rated higher than all the others, followed by US2. AU1, IE1, and US3's prosodic patterns were in a three-way tie for third place, and UK1's pattern was rated significantly lower than all the others. This, however, does not necessarily reflect an influence of these speakers' overall prosodic characteristics. As the chosen natural and TTS voice, US1 had an inherent advantage. US2 made very similar prosodic choices and also had a virtually identical pitch range, thus considerably reducing the potential for overall quality distortions, either from sparse units or from warping across an extreme range. While UK1 did indeed produce some very different prosodic patterns, she also had a higher pitch range and most importantly, as a native speaker of British English she

used different phonemes that had to be approximated in the American English TTS voice. This led to problems of segmental quality in the two methods using Matched TTS, especially in s2 (non-rhotic "morning" with a higher and more tense stressed vowel typical of UK English) and s5 ("want" and "call"). The Australian and Indian-accented speakers exhibited this problem less.

Within-subjects interactions were also significant: Method*Sentence ($F(12,360) = 29.044$, $p<0.001$), Method*Voice ($F(15,450) = 20.807$, $p<0.001$), Sentence*Voice ($F(20,600) = 16.861$, $p<0.001$), and Method*Condition*Voice ($F(60,1800) = 11.825$, $p<0.001$). The Method*Sentence interaction reflected the fact that the ratings for the four methods varied across sentences, and vice versa. There were basically two different patterns of Method*Sentence results: for s1, s3, and s4, the N+W and MT methods received the highest ratings, while DT+W and MT+W were significantly lower; for s2 and s5, the N+W method was higher than any other, followed by the DT+W method and then MT and MT+W. The difference between the two groups of sentences is best explained by the aforementioned segmental quality problems with s2 and s5.

The Method*Voice interaction was due to different ratings for the four methods across voices. The interaction effect here mainly involved voice rating differences between N+W and MT methods; ignoring the case of US1 (where unwarping natural speech was compared to TTS), N+W and MT ratings were comparable for US2 and IE1, but MT ratings were significantly lower for UK1, AU1, and US3. In all the other comparisons between methods, one method was consistently lower than the other for at least five, in one case all six voices.

There was no significant between-subjects main effect of Group (native vs. non-native listeners), although the Group*Method interaction approached significance ($F(3,90)=2.005$, $p<0.119$). Native English listeners' ratings of the N+W method were higher (3.723) than N+W ratings by non-native English listeners (3.546). For the three TTS methods, in contrast, ratings by non-native listeners were slightly (but not significantly) higher than ratings by native English speaking listeners.

5. ADDITIONAL TESTING

The perception experiment showed no conclusive results confirming a strong influence of natural prosody variation in the perception of TTS quality. Instead it appears that while listeners do perceive prosodic differences such as final tones, distribution and types of pitch accents, or phrasing structure (distribution of pauses) they do not perceive them as especially disadvantageous for overall quality perception for two reasons:

- The respective prosodic variation is not ungrammatical in that it still allows a possible interpretation (which is very easy to arrive at when no further context is given).
- Effects of prosody variation are overshadowed by the distortions present in synthetic speech (signal

processing and concatenation distortions, and pitch range and segmental discrepancies).

For this reason a small additional test was devised. The same methods and procedures were used, but the test consisted of only one test sentence:

s6: I enjoy vegetables and I also like a berry now and then.

In this particular stimulus, the DT+W method creates an inappropriate pause between “now” and “and”, for which there is no reasonable interpretation. The purpose of this additional test was to see whether this would disproportionately lower the ratings for the DT+W method, thus showing a clear influence of prosody in addition to regular synthesis-related distortions associated with this method.

Twenty-five adult listeners participated in the experiment; 15 were native speakers of English, 10 were non-native speakers of English. A Method (4) by Voice of prosodic pattern (6) repeated measures analysis of variance (ANOVA) was performed, and tests of within-subjects effects (Condition, Voice) were made for significant main effects or interactions as determined by the ANOVA. Given the result of the main test of no significant differences between native and non-native listeners, no such distinction was made in this analysis.

There were significant main effects for both Method ($F(3,72)=48.950, p<0.001$) and Voice ($F(5,120)=24.701, p<0.001$). In order from highest to lowest ratings, Methods were ranked: N+W, MT, MT+W, and DT+W. All differences among the four Methods were significant. The prosodic patterns of US1, US2 and UK1 were rated significantly higher than the other three voices; US3 and IE1 were rated significantly higher than AU1. There was also a significant Method*Voice interaction ($F(15,360)=13.298, p<0.001$), indicating that ratings of Method varied significantly among Voices, and vice versa. The additional test shows that quality ratings of DT+W decline in relation to the other methods. As Figure 2 illustrates, DT+W is the worst condition in Test 2, whereas in Test 1 it was rated slightly higher than MT+W.

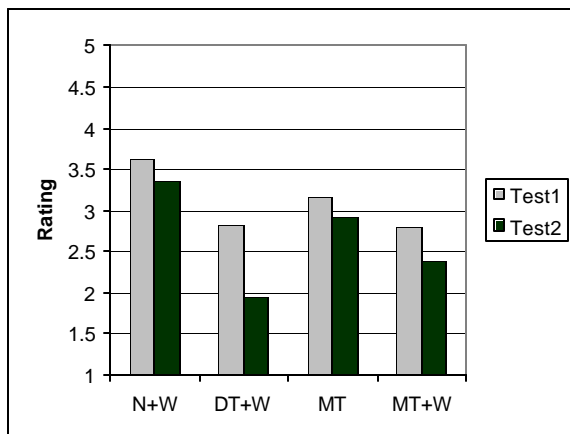


Figure 2. Comparison of ratings per method between main test (Test 1) and additional test (Test 2).

6. CONCLUSIONS

This study shows that natural prosody variations in synthesized speech do not constitute a major factor in perceived quality. Unless clearly ungrammatical phenomena are involved that preclude a meaningful interpretation, the effects of distortions associated with concatenative synthesis dominate perceptual judgments.

It must be taken into account that the methods of prosody modification used did not guarantee equal levels of distortion per sentence to allow truly fair comparisons, as they depended on the speakers' renditions of the respective sentence and their relation to the TTS voice in terms of pitch contour and pitch range at any given point within the sentence. In order to investigate the possible effect of this, an additional small-scale analysis was performed on three N+W versions of s3 (“How may I help you?”) that were judged to be without audible distortions. As the ratings (US1: 4.50; US2: 4.44; US3: 4.47) were not significantly different from each other, this may serve to support our primary conclusion. The results of the additional test suggest that only when it is very difficult to arrive at a reasonable interpretation can an effect of prosody on perceived overall quality be observed.

Future experiments should include sentences in more elaborate contexts, where a particular situation restricts the number of possible interpretations. With respect to TTS technology, this capability goes beyond models of prosody prediction into the area of text comprehension.

For the experimental synthetic speech conditions represented in this study, subtle natural prosody variations within contained rhythmic units such as Intonation Phrases do not substantially contribute to perceived quality.

REFERENCES

- [1] M. Beutnagel, A. Conkie, and A. K. Syrdal. “Diphone synthesis using unit selection,” Proceedings of ESCA/COCOSDA Third International Workshop on Speech Synthesis, pp. 185-190, 1998.
- [2] A. K. Syrdal, A. Conkie, and Y. Stylianou. “Exploration of acoustic correlates in speaker selection for concatenative synthesis,” Proceedings of ICSLP 98, pp. 2743-2746, 1998.
- [3] M. Macchi, M. Altom, D. Kahn, S. Singhal and M. Spiegel. “Intelligibility as a function of speech coding method for template-based speech synthesis,” Proceedings of Eurospeech '93, pp. 893-896, 1993
- [4] J. Pierrehumbert and J. Hirschberg, “The meaning of intonational contours in the interpretation of discourse,” in *Intentions in Communications*, P. Cohen, J. Morgan and M. Pollock, Eds., pp. 271–311. Cambridge MA: MIT Press, 1990.
- [5] J.-H. Chen, “Low Complexity Wideband Speech Coding”, Proceedings of the IEEE Workshop on Speech Coding for Telecommunications, pp. 27-28. 1995.