PROSODIC VS. SEGMENTAL CONTRIBUTIONS TO NATURALNESS IN A DIPHONE SYNTHESIZER

H.T. Bunnell, S.R Hoskins, and D. Yarrington

Speech Research Laboratory
The A. I. duPont Hospital for Children and The University of Delaware
Wilmington, DE 19803

ABSTRACT

The relative contributions of segmental versus prosodic factors to the perceived naturalness of synthetic speech was measured by transplanting prosody between natural speech and the output of a diphone synthesizer. A small corpus was created containing matched sentence pairs wherein one member of the pair was a natural utterance and the other was a synthetic utterance generated with diphone data from the same talker. Two additional sentences were formed from each sentence pair by transplanting the prosodic structure between the natural and synthetic members of each pair. In two listening experiments subjects were asked to (a) classify each sentence as "natural" or "synthetic, or (b) rate the naturalness of each sentence. Results showed that the prosodic information was more important than segmental information in both classification and ratings of naturalness.

1. INTRODUCTION

Current laboratory and commercial speech synthesizers produce highly intelligible speech, but the naturalness of synthetic speech remains a problem. The lack of naturalness in synthetic speech has been variously attributed to inappropriate modeling of the physical acoustic properties of the vocal tract [9], incorrect modeling of the articulatory and coarticulatory properties of natural speech [10], and failures in modeling the prosodic structure of natural speech [1]. Of course, it is likely that deficiencies in all these areas contribute to the perceived lack of naturalness in synthetic speech. But to what extent is each of these factors responsible for reducing the naturalness of synthetic speech?

There is very little reported data on this question; the only study we are aware of is Terken and Lemeer [11], where a listening experiment was conducted using LPC encoded speech of two levels of quality (good/poor) and two types of intonation (natural/synthetic), for both texts and individual utterances. It was found that natural intonation was always preferred for texts, but for individual utterances only preferred for good quality speech. However, this experiment was entirely based on LPC analysis and resynthesis of natural speech, so the results are not necessarily indicative of the interaction of segmental and prosodic factors in formant based, diphone, and other types of synthesizers.

In the present study, we examine the contribution of factors at the prosodic versus segmental level to the naturalness of a specific laboratory TTS system, ModelTalker [7]. ModelTalker is a data-based concatenative synthesis system for which we have developed several phoneme-to-sound modules. The version used for the present experiments was a diphone concatenation system which used automatically extracted diphones having variable context-dependent boundaries [12], and employed a PSOLA-like time domain technique for pitch and duration control.

For the purposes of this experiment, we define prosody narrowly to mean the intonation contour of a sentence and the durational pattern of the phonetic segments. There are other acoustic properties associated with prosody, such as amplitude, articulatory, and source characteristics. However, we do not currently possess good analysis/resynthesis methods for modifying these parameters. In addition, some previous studies on one particular prosodic phenomenon, focus, show that amplitude and spectral tilt play a minor role in the correct perception of focus [5][6].

We conducted two listening experiments on naturalness and synthetic speech. The experiments were designed to independently compare the naturalness of synthetic prosody to natural prosody and diphones to natural speech.

2. METHOD

2.1 Stimuli

The stimuli were drawn from a set of nonsense sentences of the form "The X is Ying the Z" (such as "The dew is leaping the bag."). The target words (X, Y, and Z) were chosen to provide closed sets of alternatives (e.g., dew, chew, Jew, …) to assess confusions among phonetically similar items. Seventy four such sentences are required to form a complete set for segmental intelligibility studies. However, for these experiments which examined naturalness rather than intelligibility, only 24 sentences were selected.

The study was made possible by the availability of the same four talkers, two male and two female, for both natural speech recording and for the recording of carrier words/phrases from which diphone data were extracted.

Speech recording and preparation procedures were similar for both natural sentences and for the utterances needed to construct diphone inventories. In particular, the talker was seated in a sound attenuated chamber before a computer monitor, keyboard, and mouse connected to a Pentium PC (located outside the chamber) running Windows 95. Talkers wore headphones with head-mounted microphone (Sennheiser HD-410), and electroglottograph (EGG) electrodes. Two channels of data were recorded at a 16 kHz sampling rate with 16-bit resolution. The first channel was the audio signal and the second channel was the output of a Glottal Enterprises EGG. An interactive program prompted talkers for speech material to record, digitized the speech, used the EGG output to locate pitch periods within all voiced regions, and aligned a phonetic transcription to the speech data using an HMM-based forced recognition algorithm. All recordings were subsequently checked by laboratory staff for errors in either the pitch tracking or the phonetic label alignment, and any errors detected were manually corrected.

In addition to the 24 sentences, each of the four talkers recorded 151 two- and three-syllable nonsense words from which small diphone inventories were extracted. These inventories were designed to provide all the diphones required to produce the synthetic semantically anomalous sentences.

From each synthetic/natural pair of sentences, two additional sentences were generated. These were produced by first computing the time-warp needed to map between the temporal structure of the natural and synthetic tokens, and then applying the time-warp pitch synchronously to map the timing and intonation of the natural sentence to those of the synthetic sentence and the timing and intonation of the synthetic to that of the natural utterance. All time-warping and pitch adjustments were done using a program which implemented the time-domain PSOLA algorithm. Thus, there were four versions of each sentence differing in the origin of their segmental and prosodic features: synthetic segments and prosody (SYNS+SYNP); synthetic segments and natural prosody (SYNS+NATP); natural segments and synthetic prosody (NATS+SYNP); and natural segments and natural prosody (NATS+NATP).

To reduce the number of sentences to a manageable size for the listening experiments, we selected eight sentence types from the 24 sentence types processed for each talker. This was done by asking four trained listeners, to rate the acceptability of all four versions of each sentence type on a five point scale. For each talker, the eight sentence types with the best acceptability score were kept. The eight sentence types chosen were not identical for each talker. In nearly all cases, sentences with very low acceptability ratings were so rated because of signal processing artifacts in the time-warping. Finally, to further reduce the contribution of signal processing alone to perceived naturalness, white noise was added to each sentence at an average SNR of +15dB.

2.2 Subjects

All subjects were undergraduate students from the University of Delaware, native speakers of American English, with no hearing difficulties. Fourteen subjects participated in the first experiment, and seventeen in the second experiment.

2.3 Procedure

For both experiments, subject were told that they would listen to a number of nonsense sentences, half from a synthetic source and half from a natural speech source. In the first experiment, subjects provided a binary classification of sentences as being synthetic or natural in origin. In the second experiment, subjects were asked to rate the naturalness of each sentence on a scale of 1 to 5, 1 being "very natural" and 5 "very synthetic". Each subject heard a total of 128 sentences (8 sentences X 2 origin conditions X 2 prosodic conditions) repeated 5 times, The sentences were presented over binaural headphones in a sound-dampened booth, with the presentation order randomized for each listener. Most of the subjects completed the listening task within 45 minutes.

2.3 Data Analysis

Repeated measures ANOVAs were performed on the data of both experiments. The dependent measure for the first experiment was the percentage of trials on which the sentence was classified as "natural". In the second experiment the dependent measure was the rating. The factorial design for the analysis was 4 (TALKER) by 2 (ORIGIN) by 2 (PROSODY) by 8 (SENTENCES) nested within subjects. In addition, omega-squared [3] was computed to provide estimates effect size for each factor.

3. RESULTS

Results from both experiments were quite similar and will be presented together. Wherever F ratios or other statistics are presented for both experiments together, results for Experiment 1 are presented first, followed by results for Experiment 2. Overall, the sentences were classified as natural 41.8% of the time, and given a mean naturalness rating of 3.36. There was a significant effect for TALKER (F[3,39]=10.38, $p<0.0001$; F[3,48]=30.11, $p<0.0001$); this is due to lower naturalness ratings of one of the male talkers (31.4% naturalness, 3.77 rating).

There were large significant effects for PROSODY in both experiments (F[1,13]=1082.56, $p<0.0001$; F[1,16]=448.19, $p<0.0001$), and Omega-squared revealed a substantial effect (Omega-squared = 0.39 and 0.16 respectively for the two experiments). On average, sentences with natural prosody, regardless of the source of the segmental information were classified as natural 71.3% of the time, and were given an average naturalness rating of 2.6 (1.0 is completely natural). By contrast, sentences with synthetic prosody were classified as natural only 12.3% of the time and received an average naturalness rating of 4.2.

There were modest significant effects for ORIGIN in both experiments (F[1,13]=207.47, $p<0.0001$ and F[1,16]=200.87 $p < 0.0001$ respectively). The Omega-squared values for this effect were 0.10 and 0.06 respectively. Sentences with natural ORIGIN (averaged over both types of PROSODY) were classified as natural 57.8% of the time, and had mean naturalness ratings of 2.8. Sentences of synthetic ORIGIN were classified as natural 25.9% of the time and had a mean naturalness rating of 3.9. It is interesting to note that these effects of ORIGIN independent of PROSODY are generally weaker than are the effects of PROSODY, independent of ORIGIN. That is, PROSODY accounts for more of the variance in these data than whether the sentences were formed from diphones or naturally uttered.

This interpretation is weakened somewhat by the significant interaction of the PROSODY and ORIGIN (F[1,13]=41.15 p < 0.0001; and F[1,16]=119.73 p < 0.001) Omega-squared values for these interaction terms were 0.04 and 0.02 respectively. Tables 1 and 2 show the means underlying this interaction for both the classification and rating data. The data in these tables suggest that the interaction was due to the much smaller effect of ORIGIN for sentences with synthetic PROSODY compared to the effect of ORIGIN for sentences with natural PROSODY.

Classification		ORIGIN	
		NATURAL	SYNTHETIC
PROSODY	NATURAL	96%	46.7%
	SYNTHETIC	19.6%	5.1%

Table 1 – Interaction of PROSODY and ORIGIN for percentage natural classifications

Rating		ORIGIN	
		NATURAL	SYNTHETIC
PROSODY	NATURAL	1.72	3.39
	SYNTHETIC	3.98	4.33

Table 2 – Interaction of PROSODY and ORIGIN for average naturalness ratings.

There were additional significant interactions from both experiments involving TALKER. Specifically, the TALKER by ORIGIN interaction (F[3,39]=7.26, p < 0.0001; F[3,48]=11.49, p< 0.0001), TALKER by PROSODY (F[3,39]=9.33, p < 0.0001; F[3,48]=23.74, p < 0.0001) and the three-way interaction of TALKER by ORIGIN by PROSODY (F[3,39]=4.44, p < 0.00001; F[3,48]=6.09, p < 0.0013). These interactions were due to the lower scores and smaller differences in the responses to sentences of the second talker. This can be clearly seen in tables 3 and 4 which show the naturalness and ratings data for each talker. The ratings and naturalness data for talker 2 are worse in almost all categories, but are most different from the other talkers in the synthetic PROSODY, natural ORIGIN condition.

EXP1	O=N
	P=N
		O=S
		P=N
			O=N
			P=S
				O=S
				P=S

T1	98.7	53.6	10.0	4.5
T2	91.2	20.5	12.1	1.3
T3	97.1	58.0	32.1	7.6
T4	96.4	54.7	24.1	6.9

Table 3 – Naturalness Judgments by Talker

EXP1	O=N
	P=N
		O=S
		P=N
			O=N
			P=S
				O=S
				P=S

T1	1.37	2.96	4.20	4.26
T2	2.15	4.08	4.24	4.61
T3	1.56	3.22	3.50	4.14
T4	1.81	3.31	3.97	4.31

Table 4 – Naturalness Ratings by Talker

4. DISCUSSION

The results show that both prosodic and segmental information affect naturalness judgments, but prosodic information appears to play a stronger role. Of course, these results are necessarily specific to the segmental and prosodic characteristics of the speech produced by the ModelTalker synthesizer. By employing two listening tasks on the same set of data, we have shown that the importance of prosody over the segmental contributions of naturalness is due to listener preference, and not an artifact of the task.

The importance of prosody can be seen most clearly in responses to the crossed conditions (synthetic PROSODY, natural ORIGIN and vice versa), where almost half of the natural PROSODY but synthetic ORIGIN sentences were judged as natural, but only 20% of the synthetic PROSODY/natural ORIGIN were considered natural. This is interesting because there are several types of speech degradation introduced at the segmental level by diphone concatenation – the synthetic ORIGIN speech was noticeably less smooth than the natural ORIGIN speech. Yet given the choice of one natural parameter and one synthetic parameter, listeners preferred natural prosody.

The experimental results agree with the findings of Terken and Lemeer [11] for the good quality speech condition. In this study, the authors found that natural or synthetic prosody did not make a difference for the poor quality speech. The data for the second talker is suggestive of this result. Although the trends in the data of Talker 2 have the same direction as the rest, the differences between the two crossed conditions are much smaller. Impressionistically, the sentences generated for this talker had the highest amount of signal processing distortion.

Since prosody overrides to a certain degree segmental information with respect to perceived naturalness, this indicates that significant improvements in speech synthesis naturalness can be obtained through better modeling of suprasegmental parameters.

5. REFERENCES

1• Akers, G. and M. Lennig (1985). Intonation in text-to speech systems: Evaluation of algorithms. J. Acoust. Soc. Amer. 77, 2157-2165

2• Allen, J., M. Hunnicutt, and D. Klatt (1987). From text to speech: the MITalk system. Cambridge University Press, Cambridge UK

3• Bagozzi R.P. (1994) (Ed.) Principles of Marketing Research. Oxford: Blackwell

4• Bezooijen, R. van and L.C.W. Pols (1990). Evaluating text-to-speech systems: Some methodological aspects. Speech Communication, 9(4), 263-270

5• Bunnell, H.T., S. Hoskins, and D. Yarrington (1997) Interactions among f0, amplitude, and duration in the perception of focus. Journal of the Acoustical Society of America, 101, 3200.

6• Bunnell, H.T., S. Hoskins, and D. Yarrington (1998) Non-traditional acoustic features of focus. Proceedings of the ICA/ASA, forthcoming.

7• Bunnell, H.T., S. Hoskins, and D. Yarrington (1998) The ModelTalker Project: Software for diphone speech synthesis and automatic diphone extraction. University of Delaware, Computer and Information Sciences Technical Report, 98-13

8• Bunnell , H.T., D. Yarrington, and K. Barner. (1994) Pitch control in diphone synthesis. In Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis. New Paltz, NY. pp. 127- 130

9• Flanagan, J.L. (1972) Speech analysis, synthesis, and perception. Springer-Verlag, New York

10• Stevens, K.N. and C.A. Bickley (1991). Constraints among parameters simplify control of Klatt formant synthesizer. J. Phon., 19, 161-174

11• Terken, J. and G. Lemeer (1988). Effects of segmental quality and intonation on quality judgements for texts and utterances

12• Yarrington, D, H.T. Bunnell, and G. Ball. (1995) Robust automatic extraction of diphones with variable boundaries. In Proceedings of Eurospeech95: Vol 3, pp. 1845-1848