International Journal of Speech Technology, 1, 7–19 (1995)

Measuring the Naturalness of Synthetic Speech

HOWARD C. NUSBAUM, ALEXANDER L. FRANCIS AND ANNE S. HENLY
Center for Computational Psychology, Committee on Cognition and Communication, 5848 South University Avenue,
The University of Chicago, Chicago, IL 60637
hcn@speech.uchicago.edu
alfr@speech.uchicago.edu
henly@ccp.uchicago.edu

Received May 18, 1995; Accepted June 16, 1995

Abstract. Even the highest quality synthetic speech generated by rule sounds unlike human speech. As the intelligibility of rule-based synthetic speech improves, and the number of applications for synthetic speech increases, the naturalness of synthetic speech will become an important factor in determining its use. In order to improve this aspect of the quality of synthetic speech it is necessary to have diagnostic tests that can measure naturalness. Currently, all of the available metrics for evaluating the acceptability of synthetic speech do not distinguish sufficiently between measuring overall acceptability (including naturalness) and simply measuring the ability of listeners to extract intelligible information from the signal. In this paper we propose a new methodology for measuring the naturalness of particular aspects of synthesized speech, independent of the intelligibility of the speech. Although naturalness is a multidimensional, subjective quality of speech, this methodology makes it possible to assess the separate contributions of prosodic, segmental, and source characteristics of the utterance. In two experiments, listeners reliably differentiated the naturalness of speech produced by two male talkers and two text-to-speech systems. Furthermore, they reliably differentiated between the two text-to-speech systems. The results of these experiments demonstrate that perception of naturalness is affected by information contained within the smallest part of speech, the glottal pulse, and by information contained within the prosodic structure of a syllable. These results show that this new methodology does provide a solid basis for measuring and diagnosing the naturalness of synthetic speech.

Keywords: synthetic speech, naturalness, intelligibility, perception

When listening to almost any kind of synthetic speech generated by rule, we are immediately aware of how unnatural it sounds. Indeed, unless a synthetic utterance has been hand-tailored to closely match the acoustic properties of a recorded natural utterance (e.g., Holmes, 1961; 1973), when there is no extraneous noise or distortion in the communication channel, synthetic speech almost always sounds different from speech produced by a human talker. From global levels of prosody to local acoustic-phonetics of spoken words, we can readily hear the source difference between natural speech and synthetic speech.

On one hand, it might seem that the voice quality of synthetic speech is less important than the intelligibility of the speech. If a listener cannot understand a message or has to work so hard to understand an utterance that performance of other tasks suffers (see Nusbaum and Pisoni, 1985), the synthetic speech will not be useful. From this perspective, the unnatural voice quality of speech seems more of an aesthetic issue than one that is important to determining the usability of synthetic speech.

On the other hand, as the intelligibility of synthetic speech improves, the naturalness of synthetic speech becomes increasingly more important. Perhaps the clearest example of this is in the area of aids for the disabled. People who have various speech or language disorders, or motor control disorders that impair speech production, can use a synthetic speech system as a vocal prosthesis (e.g., Hunnicutt, 1995). For these people, a synthetic speech system may make vocal communication possible. There is no doubt that intelligibility will be important to the use of synthetic speech in this case. However, the voice quality of the synthetic speech is also an important factor. For example, it may be more difficult to communicate using synthetic speech if the user cannot identify with the voice quality (Hunnicutt, 1995). For the talker, using computer-generated speech that sounds mechanical may be embarrassing or awkward. For the listener, it may engender attributions or beliefs about the talker that are prejudicial and inappropriate. Nobody wants to be perceived as a machine. Clearly if the speech is more human-sounding and appropriate (to the talker), communication will be more comfortable and therefore the synthetic speech will be more useful.

However, even beyond this particular situation, the perceived naturalness of synthetic speech will be an important factor in its acceptability and use in a wide range of applications. In applications involving computer interaction over a telephone, naturalness will be important. We know that many people hang up on answering machines and reject interacting with voice-mail systems. If any segment of the population has a negative response to interacting with machines, the perception that speech is produced by a computer will adversely affect the use of a system.

Indeed, there have been attempts to measure the overall acceptability of synthetic speech (see Nusbaum et al., 1984; Schmidt-Nielsen, 1995). Global measures of speech quality such as acceptability might provide a "figure of merit" that can be used to rank speech systems that take into account all relevant aspects of speech quality. However, the drawback to tests such as the Diagnostic Acceptability Measure (Schmidt-Nielsen, 1995; Voiers, 1977) and others (Nusbaum et al., 1984) is that they are highly correlated with intelligibility. This correlation means that for most intents intelligibility will provide a sufficient figure of merit. However, for the present purposes what is interesting is that acceptability measures also do indicate the listener is sensitive to other aspects of the sound of synthetic speech. Since differences on acceptability tests reflect more than intelligibility alone, it is likely that these differences reflect the relative naturalness of synthetic speech. Unfortunately, because of the confounding of intelligibility in these tests, they do not give any separate information about naturalness.

If naturalness is important to the acceptability of synthetic speech, it is important to measure it. Aside from the goal of ranking systems on naturalness, improvements in the naturalness of synthetic speech systems will depend on accurate measures. While subjective impressions of naturalness may be useful overall to developers seeking to improve their systems, since these impressions are psychologically confounded with intelligibility it will be difficult to diagnose specific problems of naturalness from these impressions. Although there has been some question raised as to whether or not diagnostic tests of intelligibility have had a substantial impact on intelligibility improvements in text-to-speech systems (Pols and van Bezooijen, 1991), there are cases in which specific acoustic-phonetic diagnoses have assisted in improving aspects of intelligibility (cf. Logan et al., 1989). Simply on logical grounds however, it seems quite plausible that if our overall impressions of synthetic speech are confounded with intelligibility, diagnostic tests that are less influenced by intelligibility may be of greater use in improving the naturalness of synthetic speech.

The Problem of Describing Naturalness

There is no extant objective definition of naturalness that we are aware of—it is a voice quality that is purely subjective. Thus there is no filter or signal processing algorithm that we can apply to a sample of speech that will yield a measure of naturalness. However, we can specify analytically some of the factors that may influence the perception of naturalness. In principle, many of these factors would be similar to the factors that influence the perception that speech is "accented" when produced by a non-native speaker of the language (cf. Flege, 1988).

Synthetic speech differs from natural speech in prosodic and segmental structure and source characteristics. Thus each of these may contribute in part to the perception of synthetic speech as unnatural. Given that segmental duration and timing, intonation, and amplitude variation are under the control of rules, the patterning of these sources of information may show less variability than human speech and may be wrong or uncoordinated. Prosody in human sentences is extremely complex (e.g. Bollinger, 1989; Cooper and Paccia-Cooper, 1980; Cooper and Sorenson, 1981); the rules that are used to govern these factors in synthetic speech are limited by our scientific understanding and probably oversimplify the actual use of prosody in natural speech. In part, this oversimplification, together with actual errors in the rules, may give rise to the perception that synthetic speech is unnatural. Even in the case of research specifically directed at improving prosodic characteristics, such as segmental durations (e.g., Campbell and Isard, 1991; Klatt, 1976; Syrdal, 1989), it is clear that there is a large difference between synthetic speech and natural speech.

Similarly, at the segmental level, there are many opportunities for oversimplification and error in the rules of a text-to-speech system (e.g., Fant, 1991; Nusbaum and Pisoni, 1985). These opportunities exist both at the level of acoustic-phonetic rules and at the level of phonological rules. If the patterning of phonological segments is overly simple or contains errors, and if the use of acoustic cues in conveying phonetic segments is overly simple or contains errors, listeners will perceive this. While it is likely that many of these affect intelligibility, it seems plausible that they may affect naturalness at least as much. For example, synthesizers that differ in the degree to which transition duration and voice-onset time (VOT) vary with place of articulation and surrounding vowels will certainly differ in intelligibility since these are cues exploited by natural talkers (e.g., Lisker and Abramson, 1967). However, listeners may also perceive the lack of appropriate covariation as unnatural. Although there has been some psychophysical work regarding the sensitivity of listeners to specific cues such as changes in formant frequency and amplitude (e.g., Flanagan, 1955; 1957), and some research into the covariation of cues such as $F1$ extent and silence duration in stop consonant perception (e.g., Best et al., 1981), this work is insufficient to produce natural sounding synthetic speech.

Both segmental and suprasegmental simplifications and errors occur due to problems in the rules of a text-to-speech system. Even when rules are not involved, synthetic speech may be unnatural. In essence, if the source characteristics of synthetic speech are distinct from a human glottal waveform, it seems likely that listeners will perceive this difference. Other aspects of the glottal source affect perception of voice qualities such as creaky or laryngealized or male vs. female (see Klatt and Klatt, 1990; Laver, 1980). Thus it should not be surprising that one aspect of naturalness should be determined by source characteristics. Indeed, the source characteristics of human speech are extremely complex (e.g., Laver, 1980) and in the past many synthesizers treat the glottal waveform as a simple pulse train. Even as the modeling of source characteristics has become more sophisticated (e.g., Klatt and Klatt, 1990), it seems likely that listeners may be sensitive to the differences between a synthetic source and a natural source. For example, using hand-synthesized speech, Carrell (1984) has shown that listeners can reliably differentiate different glottal waveforms across the same vocal tract. This perceptual sensitivity has driven much of the research on improving the source characteristics of synthetic speech.

Measuring Naturalness

If we start from the position that naturalness is unlike a voice quality such as creaky voice for which there is roughly a single dimension that can be examined (e.g., Klatt and Klatt, 1990; Laver, 1980), then the measurement of naturalness becomes a serious problem. While it may be possible to specify the acoustic characteristics of a creaky voice or a breathy voice and therefore measure these characteristics in speech, there are many ways in which speech may be heard as unnatural. Furthermore, a single measure of naturalness might be globally informative in the same way as a measure of acceptability might be, however it would not be very diagnostic of the specific problem. As a result, a global measure would be of little value to researchers and developers trying to improve the quality of synthetic speech.

In this respect then, the problem of measuring naturalness is similar to the problem of measuring intelligibility. For example, intelligibility varies with the speech rate and intonation of sentences produced by a text-to-speech system and it may vary differently across different synthesizers (see Slowiaczek and Nusbaum, 1985). Intelligibility varies as a function of the experience listeners have with speech produced by a particular synthesizer. The more listeners hear synthetic speech, the intelligibility of the speech significantly improves (Schwab et al., 1985), suggesting that they are shifting attention away from misleading or inappropriate cues (Lee and Nusbaum, 1989). Measured segmental intelligibility varies as a function of the linguistic complexity of the materials and the complexity and demands of the intelligibility task (see Nusbaum and Pisoni, 1985). Even when segmental intelligibility is held roughly constant, there will be differences in how listeners comprehend the output of different synthesizers (see Ralston at al., 1995). Thus, it has been recommended that intelligibility of synthetic speech be assessed using tests that are specifically designed to satisfy the goals of the assessment (Nusbaum and Pisoni, 1985). In other words, generic tests will not be sufficient—tests must be tailored in specific ways to address specific questions. For example, assessing segmental intelligibility using the Modified Rhyme Test (House, et al., 1965) will be sufficient to compare text-to-speech systems when there are substantial differences (see Logan et al., 1989). However, it is conceivable that several systems might be equally intelligible for monosyllabic words which are used as materials in the MRT, but they might differ substantially on polysyllabic words because of differences in implementing stress and phonological rules. Also, lexical knowledge aids recognition of polysyllabic words more so than monosyllabic words (e.g., Pisoni et al., 1985).

Given these issues, how do we measure naturalness? First it is important to eliminate, as much as possible, the contribution of intelligibility to the measurement of naturalness. At the present time, intelligibility differences between natural and synthetic speech are still sufficiently large to affect perception of naturalness. Second, it is important to develop tests that target specific aspects of naturalness rather than provide global ratings. Thus, different tests are needed to assess naturalness of source characteristics and naturalness of prosody.

Experiment 1: Naturalness of Source Characteristics

The purpose of the first experiment was to attempt to assess the contributions of synthesizer source characteristics to the perception of naturalness. In order to do this it was important to reduce, as much as possible, contributions of segmental structure and prosody to the perception of naturalness of speech. Our view is that even at the level of an individual glottal pulse there is a difference between natural and synthetic speech that could be perceptible to a listener. Research on improving the source characteristics of synthetic speech has certainly indicated that such differences exist and, in terms of synthesizing differences between male and female voices, such distinctions can be perceived by listeners (e.g., Klatt and Klatt, 1990). Given the complexity and variability of natural source characteristics (e.g., Laver, 1980), it seems reasonable that even the smallest acoustic event in speech—the glottal pulse—should be perceptible as natural or unnatural.

However, we did not want to give listeners glottal waveforms extracted by inverse filtering of utterances because those signals would be extremely unnatural to begin with. Our approach then was to take a single glottal pulse from a production of a sustained vowel and iterate and concatenate the pulse to produce a new sustained vowel. By iterating pitch pulses taken from a sustained vowel, we eliminate the effects of prosody. By focusing on maximally discriminable single vowels isolated from context that are known, a priori, to the listeners, we can eliminate much of the effects of intelligibility on perception. The primary drawbacks to this approach of iterating a single pitch pulse taken from a vowel are: First, this procedure makes natural speech sound like synthetic speech and therefore reduces the range of judgments possible. Second, this is not a pure measure of source characteristics since the glottal pulse is still convolved with the resonators of the vocal tract. Thus different vowels, which have different pole-zero patterns in the transfer function will reveal different aspects of the glottal waveform (e.g., see Klatt and Klatt, 1990).

Although this method might, in principle, reveal differences in the shape of the glottal pulse that listeners perceive as differentiating synthetic and natural speech, there is one clear limitation. By iterating a single glottal pulse to form a sustained vowel, we are eliminating one aspect of source characteristics that could be important to the perception of naturalness. This approach eliminates any variability between glottal pulses which listeners might perceive as a characteristic of naturalness. The perception of variability between glottal pulses therefore needs to be examined separately. We constructed a second set of stimuli based on a sample of five successive glottal pulses that were iterated, in an attempt to measure the contribution of this variability.

The primary task for subjects was to listen to a sustained vowel and decide whether it was produced by a human or a computer. Subjects were told that all the speech had been processed by a computer and so even speech produced by a human would sound somewhat unnatural. We measured the speed and accuracy of their decisions.

Method

Subjects. The subjects were six students at the University of Chicago. All subjects were right-handed, native English speakers with no history of speech or language disorder. None of the subjects reported any prior experience listening to synthetic speech produced by a text-to-speech system. Subjects were paid $6 for their participation in the experiment.

Stimuli. We constructed two stimulus sets for this experiment. Both sets were constructed using a waveform editor to iterate glottal pulses extracted from vowels to produce a 1 second long vowel. In one set, the test stimuli were constructed from iterations of a single glottal pulse. In the second set, five successive glottal pulses were iterated as a group.

Two male talkers produced the sustained vowels /a/, /i/, and /u/ in the carrier sentence, "Say the vowel_______ please" where the blank was filled in by one of the vowels. The same three sentences were synthesized using a DECtalk PC DOS V4.0 text-to-speech system (set to the Paul voice) and a Votrax Type-'N-Talk. All of the speech was digitized at 16-bit resolution at 10 kHz after low-pass filtering at 4.8 kHz. The isolated vowels were edited out of the carrier sentence and stored in separate waveform files.

The glottal pulses were extracted from the center of each vowel using a digital waveform editor. Glottal pulses were excised at the most stable portion of the vowel and cut at a zero crossing. The one-pulse samples were then copied into a buffer until a one-second vowel was created. Thus, there were three one-pulse vowels created for each of two human talkers and two text-to-speech systems. The same procedure was carried out with the five-pulse samples to create another three one-second vowels for each of the talkers. The amplitudes of all the vowels were then digitally scaled to the lowest RMS amplitude vowel in the set which was 45 dB.

Waveform files were converted to analog form in realtime using a 12-bit D/A converter at a 10 kHz sampling rate. The signals were low-pass filtered at 4.8 kHz. Stimuli were played over Sennheiser HD430 headphones at about 65 dB SPL.

Procedure. In each experimental session, subjects listened to the stimuli constructed from one-glottal pulse first, and then they listened to the stimuli constructed from the sequence of five glottal pulses. Within each half of the session, subjects first received a set of familiarization trials. In each set of familiarization trials, subjects heard all the vowels produced in the order / a / followed by / i / followed by / u /. Each vowel was presented accompanied by a text message identifying the vowel. This text was shown on the computer display screen in front of each subject. One of the four voices was selected at random and each of the three vowels was played in sequence. Then the next voice was selected at random and each of the vowels was presented. Each voice was presented once, selected at random, and then the process was repeated. Thus subjects heard each voice produce each vowel twice. During this block subjects were told only to listen to the vowels—no response was required. The familiarization block allowed subjects to learn how each voice produced each of the vowels.

Within each half of the session, following the familiarization block, subjects were give a block of test trials. In the test trials, each of the three vowels produced by each of the four voices (two natural and two synthetic) were presented 16 times. The ordering of stimuli was random.

The subjects were instructed to listen to each vowel and decide as quickly and accurately as possible whether it was produced by a human or by a computer. Two response keys on a keyboard were marked H and C and the labels HUMAN and COMPUTER appeared on the display screen. Subjects were encouraged to guess if they were unsure. Response times were measured with msec accuracy.

Results and Discussion

There are two ways in which subjects' responses could be meaningfully scored and analyzed. We could examine the *accuracy* of subjects, naturalness decision. In this case, a correct response for a human voice is to call it human and a correct response for synthetic speech is to classify it as produced by a computer. This would be the way we might score responses if we wanted to understand how accurate listeners are in source classification.

However, for our present purposes, we are more interested in measuring naturalness. In this case, we are interested in finding a method that allows us to rank order speech according to how natural the speech sounds. This means that speech produced by humans should always be assigned high values on this scale and that speech produced by text-to-speech systems should be assigned lower values on this scale with differences in perceived naturalness among synthesizers resulting in different scale values. Thus, for this goal, it is important to score the probability of classifying each sample of speech as human. The probability of calling a particular sample of speech "human" represents the possible scale for naturalness.

Figure 1 shows the probability of classifying each of the three one-glottal-pulse vowels / a /, / i /, and / u / as human when produced by the two human talkers Male 1 and Male 2, and by the DECtalk and Votrax text-to-speech systems. Figure 2 shows the probability of human judgments for the five-glottal-pulse vowels. Remember that all the stimuli were actually one second in duration. The difference between these stimulus sets is whether the one-second vowels were constructed by repeatedly concatenating a single glottal pulse or a sequence of five glottal pulses. One pulse repeated provides a snapshot of the naturalness of source characteristics. We had hoped that five pulses taken in

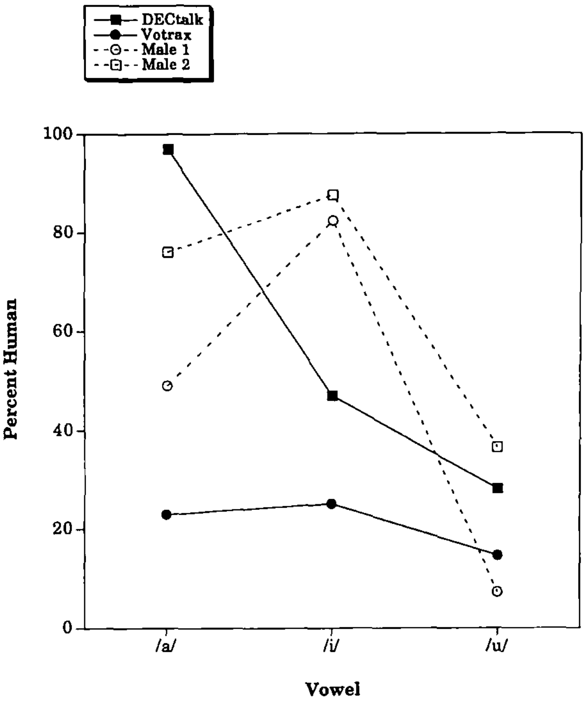

Fig. 1. Mean percent human judgments for / a /, / i /, and / u / vowels that were constructed by iterating a single glottal pulse taken from each of two male talkers and two text-to-speech systems.

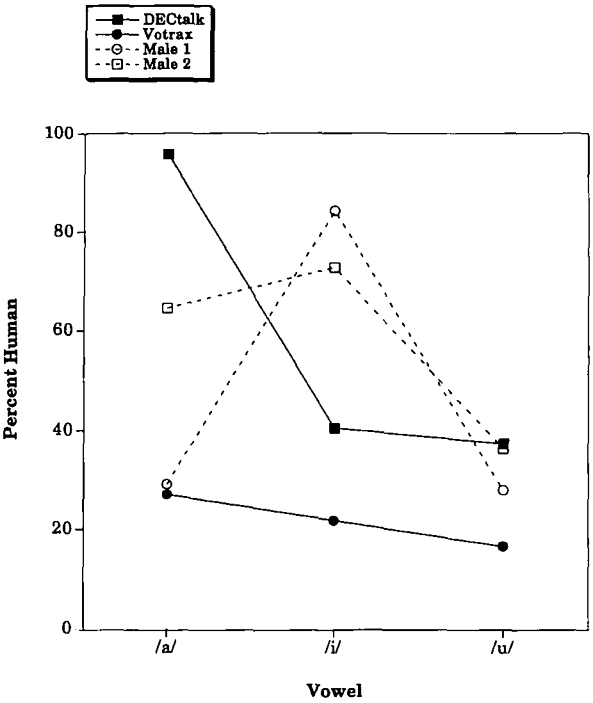

Fig. 2. Mean percent human judgments for / a /, / i /, and / u / vowels that were constructed by iterating a sequence of five successive glottal pulses taken from each of two male talkers and two text-to-speech systems.

succession would provide some information to listeners about variability of those source characteristics. Unfortunately, since the overall pattern of human judgments appears to be the same for both the one-pulse and five-pulse stimuli it is impossible to determine from this study whether variability in glottal pulses affects perceived naturalness.

As can be seen in Figs. 1 and 2, the general patterns of naturalness judgments differ across the three vowels and differs for each of the different voices. However the overall pattern is the same for both stimulus sets. An analysis of variance showed no reliable differences in the classification performance for the one-pulse and five-pulse stimuli. As a result, the data from both sets were combined into a single analysis examining the effects of the type of vowel (/ a / vs. / i / vs. / u /) and voice (two human talkers and two text-to-speech systems) on naturalness judgments.

First, a significant effect of vowel identity on the classification responses was observed, $F(2, 10) = 25.38$, $p < .01$. As can be seen in both Figs. 1 and 2, there is a tendency for / u / stimuli to be classified overall as less natural (a .26 probability of being judged as produced by a human) than either / a / (.58) or / i / vowels (.58). It is possible that / u / constructed from repetitions of a glottal pulse sounds more unnatural than / a / or / i /, possibly because, when produced as an isolated vowel, there is a tendancy for more formant movement in / u /. The lack of formant movement in the pulse-iterated / u / may therefore sound more unnatural than for / i /. Also, it has been claimed that the static vocal tract (i.e., formant) specification of / i / is unique relative to other vowels so formant movement may be less important to its definition (e.g., Lieberman and Crelin, 1971; Liberman et al., 1972).

We also found a significant difference overall in the naturalness judgments for the different voices, $F(3, 15) = 4.83$, $p < .02$. Overall, averaging across the different vowels, speech produced by the two human talkers was classified as natural with the same probability (.5 for one talker and .6 for the other); DECtalk generated speech was also classified as natural with the same overall probability of .6. None of these classification probabilities were reliably different from each other. However, Votrax was classified significantly less natural overall at .2 than the two human talkers, $F(1, 15) = 10.56$, $p < .01$. Also Votrax speech was judged reliably less natural than DECtalk speech, $F(1, 15) = 9.48$, $p < .01$.

Finally, as can be seen in Figs. 1 and 2, there was an interaction between voice and vowel so that the basic pattern of naturalness judgments differed for each voice across the vowels, $F(6, 30) = 9.47, p < .01$. For example, listeners consistently classify speech produced by Votrax as computer-generated and there is little variation in these judgments across the vowels. However this pattern contrasts sharply with judgments of DECtalk speech which span the entire range of naturalness depending on the identity of the vowel. While the DECtalk speech is also judged unnatural for / u / and / i /, listeners perceive it to be extremely natural and human sounding for the vowel / a /, in fact more so than the human speech, $F(1, 30) = 22.55, p < .01$. Similarly, patterns of naturalness judgments differ across the three vowels for speech produced by the two human talkers. The pattern for the two human talkers is generally similar although there are some differences between the talkers (see naturalness judgments for the vowel / a / in Figs. 1 and 2).

Overall though, it appears that naturalness judgments for the vowel / u / are more biased by the vowel itself than the differences among the voices. All four voices are heard as unnatural for / u /. One possible reason for this is the lack of formant movement in / u / may sound unnatural in these vowels constructed from iterated pitch pulses since formant movement may be more expected as part of its identity. Thus, / u / is probably not a good vowel for measuring the naturalness of synthetic speech using the iterated glottal pulse method since it does not differentiate natural from computer-generated speech and does not differentiate among speech produced by the different synthesizers.

Similarly, / a / does not provide a good measure of naturalness. Although for this vowel we found that naturalness judgments are reliably different for human speech compared to synthetic speech and are reliably different for the two synthesizers, the extremely high naturalness ratings for DECtalk were unexpected. For some reason, the DECtalk iterated-glottal-pulse / a / sounds more human than the speech produced by humans. While this indicates that the source function works well for DECtalk for this particular vowel, given the formant resonances, it also indicates that / a / will not be a good diagnostic stimulus for assessing naturalness. Although, it is important to note that for particularly unnatural sounding speech, such as that produced by Votrax, / a / is reliably diagnostic.

The pattern of data suggests instead that / i / does provide a good index of naturalness, with all the properties we would like in a naturalness scale: First, speech produced by the human talkers is classified as natural more often than DECtalk-produced speech, $F(1, 30) = 18.77, p < .01$, and is classified as natural more often than Votrax-generated speech, $F(1, 30) = 44.19, p < .01$. Second, speech produced by DECtalk was judged more natural than speech produced by Votrax, $F(1, 30) = 4.02, p < .05$. This suggests that it is possible to measure the naturalness of the glottal source using the present method.

There are several conclusions that can be drawn from the present results. First, the present results demonstrate that it is possible to develop a test that assesses naturalness at the most microscopic level of the speech signal—the characteristics of the glottal source. Our results clearly indicate that iterating a glottal pulse from the center of an / i / vowel provides diagnostic information about the naturalness of the speech. We believe that this naturalness judgment is primarily influenced by the source characteristics of the speech rather than intelligibility. Since the set of vowels used in the present study were all clearly intelligible, known to the listeners ahead of time, and the listeners were familiar with the specific sound of these vowels, this test is not influenced by the way intelligibility normally covaries with naturalness. It is, of course possible, that the listeners are judging vowel quality rather than naturalness. However, there are no apparent differences in vowel quality—rather there are differences in the sound of the source characteristics. Thus, we believe that listeners are judging source characteristics in the present study. Among the vowels we tested, / i / may be more transparent for source characteristics because of the low $F1$ and high $F2$ capturing both low- and high-frequency components of the glottal source. In particular, the higher frequency components of the glottal source are important to determining the shape of the glottal waveform, and those components are more damped with the lower $F2$ of / a / and / u /.

Second, it is also clear that the naturalness of the glottal source characteristics of high-quality synthetic speech such as that produced by DECtalk has moved into a range that is close to human speech. Across the different vowels, Votrax-produced speech was always perceived as unnatural. However, DECtalk speech was perceived as varying throughout a range of naturalness that is spanned by human speech. Thus while substantial gains in naturalness may be achieved by improving the source characteristics of lower cost text-to-speech systems, at the high end the impact of research on improving source characteristics of synthetic speech has clearly been substantial (Klatt and Klatt, 1990).

Still the use of a diagnostic test such as the present one based on /i/ may guide further improvements at this level of the naturalness of computer-generated speech.

Finally, we believe there is a contribution of glottal pulse variability to the perception of naturalness in synthetic speech (cf. Fant, 1991; Klatt and Klatt, 1990; Laver, 1980). Although we attempted to measure this contribution by comparing vowels synthesized by iterating a single glottal pulse with vowels synthesized by iterating a set of five successive glottal pulses, we found no systematic difference in the perceived naturalness of these two stimulus sets. We do not believe this means that there is no effect of glottal pulse variability. Rather, we believe that either the five-pulse sample is not a sufficient sample of the glottal pulse variability or the repetition of the five-pulse set destroys any contribution this small amount of variability might make to the perception of source naturalness. Thus, it appears that in order to measure the effects of glottal pulse variability on perception of naturalness, longer samples of speech may be needed.

Experiment 2: Naturalness of Lexical Prosody

As we noted previously, the perceived naturalness of speech may result from a number of different acoustic properties, from the level of the source characteristics of the speech, through the segmental structure of the speech, to the prosodic properties of utterances. Our first experiment demonstrated that even at the most basic level of speech, the glottal pulse, there is information that affects the perception of naturalness. Clearly there is also evidence that segmental intelligibility is used as an indicator of naturalness for listeners as well (Nusbaum et al., 1984; Schmidt-Nielsen, 1995). It seems reasonable that listeners can use the acoustic-phonetic structure of speech as information that speech is not produced by a human talker. Furthermore, it is apparent, even from subjective listening, that sentential prosody contributes substantially to the perception of naturalness (Fant, 1991; Kohler, 1991; Syrdal, 1989). However, there is apparently little evidence about the role of lexical prosody in the perception of naturalness. Certainly, in linguistic terms, we know about the contributions of lexical structure to prosody (e.g., Selkirk, 1986). But can listeners use this kind of word-level prosody reliably as a basis for the perception of naturalness? If lexical prosody contributes to naturalness perception, this will represent an important area to target for research and development of improvements in text-to-speech systems.

The present study was carried out to measure the sensitivity of listeners to differences in lexical prosody produced by humans and text-to-speech systems. Again, as in our previous experiment, our goal was to eliminate the influence of intelligibility differences and focus on the perception of prosody. To do this we low-pass filtered spoken words to eliminate as much segmental information as possible, leaving only the prosodic structure. Again, as in our first study, this kind of manipulation produces an unnatural speech signal which could mask any of the naturalness differences between human- and computer-generated speech.

Undoubtedly, there are differences in intonation, amplitude, and segmental durations in words produced by humans and text-to-speech systems. But there is no a priori reason to believe that the prosody produced by a speech synthesizer by rules is outside the range of prosody produced by humans. When low-pass filtered, the perception of these signals could be so dominated by the unnatural sound of the speech that listeners may not perceive the differences among the natural and computer talkers. On the other hand, if naturalness differences are perceived reliably in spite of this signal processing, this would be a strong indication of the perceptual salience and importance of this information.

The primary goal of the present experiment, then, was to determine if listeners could reliably judge whether low-pass filtered words were produced by a human or a computer. The second goal was to determine whether or not these judgments would also differentiate between utterances produced by text-to-speech systems differing in naturalness. In our first experiment, naturalness judgments for the vowel /i/ not only distinguished between human and computer talkers, it also distinguished between two synthesizers differing in naturalness. This provides the basis for a naturalness scale that can be used to compare text-to-speech systems as well as diagnose changes in a text-to-speech system that should affect the naturalness of its speech. Similarly, in our second experiment, our goal was to determine whether lexical prosody provides a basis for measuring the degree of naturalness of synthetic speech.

As in the first experiment, the subjects, task was to listen to a stimulus and decided whether it was produced by a human or a computer. All the words were low-pass filtered to remove all segmental detail. Subjects listened to short, monosyllabic words, disyllabic words, and polysyllabic words. We varied word length to determine whether longer words would provide more reliable information about the naturalness of the speech.

Method

Subjects. The subjects were 17 undergraduate and graduate students at the University of Chicago. All subjects were native English speakers with no history of speech or language disorder. None of the subjects reported any prior experience listening to synthetic speech produced by a text-to-speech system. Subjects were paid $8 for their participation in the experiment.

Stimuli. The stimuli for this experiment consisted of a set of 11 monosyllabic words, 27 disyllabic words, and 33 polysyllabic words (three- and four-syllable words). These words were spoken by two male talkers and the DECtalk and Votrax text-to-speech systems described in Experiment 1. For the synthetic speech, the words were entered into text files and synthesized using normal spelling (i.e., there was no correction for pronunciation errors) with pauses between words. The speech produced by the two males was digitized with 12-bit resolution at 10 kHz after low-pass filtering at 4.8 kHz and stored in digital waveform files; the synthetic speech was digitized with 16-bit resolution at 10 kHz and low-pass filtered at 4.8 kHz. The individual words were edited and saved into separate waveform files.

All the waveform files were low-pass filtered at 200 Hz to eliminate segmental information. The amplitudes of all the filtered waveforms were digitally adjusted to be the same RMS level of 71 dB. Waveform files were converted to analog form in realtime using a 12-bit D/A converter at a 10 kHz sampling rate. The signals were low-pass filtered at 4.8 kHz. Stimuli were played over Sennheiser HD430 headphones at about 68 dB SPL.

Procedure. In each experimental session, subjects participated in three blocks of trials. Each of these blocks was preceded by a small set of familiarization trials. This familiarization consisted of two repetitions of one word spoken by each of the four voices (two natural, DECtalk, and Votrax). Thus, in this familiarization the same word was spoken twice by each of the human talkers and each of the synthesizers. Familiarization used a word that was not presented in the testing block but had the same length in syllables as the test stimuli. One block consisted of 2 repetitions of the monosyllable words, a second block consisted of a single repetition of each of the disyllable words, and the third block consisted of a single repetition of each of the polysyllabic words. The order of presentation of the blocks was varied across subjects. In each block, subjects were told they would hear speech produced by a mix of human talkers and text-to-speech systems. They were told that the speech had been low-pass filtered so that they would not be able to identify the word that was spoken, but that this should not affect their ability to determine whether the word was produced by a human or a synthesizer. Subjects were told to press either of two response keys labeled Human or Computer as quickly and accurately as possible. They were told to guess if they were not sure whether the speech was produced by a human or a synthesizer.

Results and Discussion

As in the first experiment, subjects, responses could have been scored for accuracy (i.e., classifying human speech as human and synthetic speech as synthetic) or for the probability of classifying a speech sample as produced by a human. Again, we used the latter scoring method since our goal in this experiment is to develop a scale of naturalness uncontaminated by intelligibility. Thus, for these analyses we will treat the probability of classifying speech as produced by a human as a measure of naturalness. When the probability is high, this means the speech sounds extremely natural to the listener and when the probability is low, the speech sounds unnatural.

Figure 3 shows the mean probability of classifying the filtered words as human for the two male talkers and for DECtalk and Votrax for each of the three types of words (one syllable, two syllables, or polysyllable). As can be seen in Fig. 3, there was a significant effect of talker on the probability of judging the speech to be produced by a human, $F(3, 48) = 45.66, p < .01$. Although there was no difference in naturalness judgments between the two male talkers, $F(1, 48) = .38$, n.s., there was a difference in naturalness judgments between the two human talkers and DECtalk, $F(1, 48) = 52.31, p < .01$, and between DECtalk and Votrax, $F(1, 48) = 11.02, p < .01$. Across all words, listeners judged human speech the most natural (.75 for one talker and .71 for the other), and they judged DECtalk next most natural (.34) and Votrax least natural (.14).

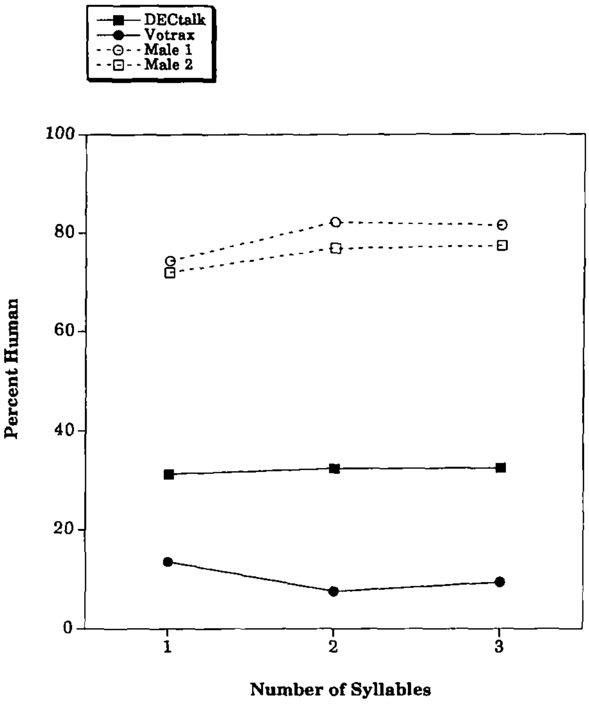

Fig. 3. Mean percent human judgments for monosyllable, disyllable, and polysyllable words produced by two male talkers and two text-to-speech systems.

Also, as can be seen in Fig. 3, there was little, if any, effect of the different word lengths on naturalness, $F(2, 32) = .91$, n.s. Naturalness did not vary reliably across the different word lengths. Furthermore, there was no interaction between talker and word length, $F(6, 96) = .83$, n.s., indicating that naturalness judgments were unaffected by word length for all the voices.

These results demonstrate that listeners do perceive the differences in prosody between natural and synthetic speech, even for single words. Indeed, even for single syllables listeners can accurately and reliably differentiate between natural and synthetic speech. Moreover, listeners can reliably judge naturalness in spite of the fact that the speech has been low-pass filtered to eliminate almost all segmental detail.

Furthermore, the probability of classifying speech as natural or synthetic does have the properties we want in a naturalness scale. First, listeners do not differentially classify the speech produced by humans—natural speech is accurately and consistently classified. Second, low-quality synthetic speech is also accurately and consistently classified as synthetic. Finally, high-quality synthetic speech is rated higher than the low-quality synthetic speech, but substantially and reliably lower than natural speech. Given that this method distinguishes reliably different levels of naturalness for synthetic speech, and that high-quality synthetic speech is substantially lower on this scale than natural speech, the present method succeeds in providing a diagnostic measure of naturalness unconfounded by the intelligibility of the speech.

General Discussion

As the quality of synthetic speech improves, the need for new diagnostic tests will increase. First, the primary focus of most diagnostic tests is segmental intelligibility, measured in relatively narrow ways (see Schmidt-Nielsen, 1995). For example, older tests of segmental intelligibility such as the Modified Rhyme Test (House et al., 1965) may no longer diagnose problems because of insufficient sensitivity. Monosyllabic test stimuli and a limited-response-set task may present too little contextual and cognitive variability as intelligibility increases (e.g., see Nusbaum and Pisoni, 1985). Some text-to-speech systems (e.g., DECtalk) already come close to the segmental intelligibility of natural speech, at least when measured using the MRT (Logan et al., 1989). There is a need for tests that measure segmental perception across a wider range of linguistic contexts and tasks.

Second, there is a need to go beyond measures of segmental intelligibility to measure word perception, sentence comprehension (Ralston et al., 1995), and naturalness. The quality of synthetic speech has long been limited by segmental intelligibility since recognizing segmental structure in synthetic speech presented the largest problem for listeners. However, as segmental quality improves there are many areas that need to be addressed in assessing the quality of synthetic speech. For example, it is important to measure how well a text-to-speech system uses prosody to aid listeners, segmentation of the speech stream into individual words, to convey syntactic, semantic, and pragmatic information, and to sound like a human talker. There are few if any standard methods for measuring any of these aspects of speech processing.

Finally, as synthetic speech quality improves, it will be increasingly difficult for developers to rely on analytic methods and their own subjective evaluations of synthetic speech. Although Pols and Bezooijen (1991) have argued that, to date, diagnostic tests have had little direct measurable impact on the development of text-to-speech systems, the gross nature of problems in synthetic speech have made it relatively straightforward to diagnose and remedy many of the problems in segmental intelligibility. Our previous research on perceptual learning (Schwab et al., 1985) demonstrates that experience listening to a particular synthetic speech system changes the perceptual processing of that speech. Thus the evaluation of synthetic speech by the experienced listener is altered and obscured somewhat by that experience, compared to the naive listener (Nusbaum and Pisoni, 1985). As the problems in synthetic speech become more subtle, it will become easier to pinpoint these problems with naive listeners who are more sensitive to the differences between a particular text-to-speech system and natural speech.

Since we cannot identify analytically all the acoustic properties that are used by listeners in understanding spoken language, the human listener becomes the standard of measurement for synthetic speech quality. This becomes particularly true when we consider the naturalness of synthetic speech. While there are numerous algorithms for word recognition that, in principle, might be used to measure intelligibility, there are no existing comparable measures for naturalness. Judging a particular utterance to be produced by a human rather than a computer is a subtle perceptual task that depends on a number of diverse acoustic characteristics of the way speech is produced. Thus, in order to improve the naturalness of synthetic speech, the use of diagnostic tests may be particularly important.

However, for diagnostic tests to be useful in improving synthetic speech, they must satisfy several constraints. Diagnostic tests must be reliable, which means that they produce consistent results when carried out under comparable conditions. This is a prerequisite for any standardized test so that it can be administered by different laboratories and still allow a meaningful comparison of the results. Also, diagnostic tests must be valid. This means they must measure what they are intended to measure. For example, tests for naturalness need to control for the influence of intelligibility. Furthermore, tests must be sufficiently sensitive to measure small differences among different text-to-speech systems so that they can accurately register changes made during the development process. In addition, tests must focus on specific problems. General tests that simply indicate that speech produced by one text-to-speech system is more natural than speech produced by a different system may be useful for deciding which system is appropriate for a particular application. But this will not provide precise diagnostic information regarding the specific deficits in a text-to-speech system that need to be remediated. In particular, this will be important in attempting to diagnose problems in the naturalness of synthetic speech.

Some tests of segmental intelligibility such as the MRT (House et al., 1965; Logan et al., 1989) and the DRT (Voiers, 1983) can provide specific diagnostic information about which phonetic segments are less intelligible. This can assist developers in improving the phonetic implementation rules for a text-to-speech system. Similarly, there is a need to pinpoint precisely the areas of a text-to-speech system that need to be improved in naturalness.

Previous tests of acceptability and naturalness (e.g., Nusbaum et al., 1984; Schmidt-Nielsen, 1995) have been too general to be diagnostically useful. Furthermore, the confounds with intelligibility in these tests make it difficult to diagnose naturalness of synthetic speech in any meaningful way. The present experiments provide two new methods for measuring naturalness that offer a solution to this problem. We have demonstrated in both experiments that listeners can reliably judge whether speech was produced by a human or a computer even when they are not listening to "complete" intact utterances. This allows us to eliminate some sources of information and focus listeners, attention on the remaining information. Thus, it is possible to reduce the contribution of segmental intelligibility to these judgments. In addition, it allows us to construct tests that focus on particular aspects of speech that may differentially contribute to the perception of naturalness such as the source characteristics of the glottal waveform or lexical prosody.

Furthermore, in our experiments we found that the probability of classifying an utterance as produced by a human is related to the naturalness of the speech. In other words, this task does not just distinguish the broad categories of natural and synthetic speech. It also distinguishes among different levels of naturalness found for different text-to-speech systems. This means that we can use this task as the basis for developing standardized diagnostic tests of naturalness.

In the first experiment, we demonstrated that listeners can reliably distinguish the naturalness of human talkers from DECtalk and from Votrax and can reliably distinguish the naturalness of the two synthesizers for productions of /i/. Although there has been attention paid to the importance of source characteristics in accurately modeling human talkers (e.g., Fant, 1991; Klatt and Klatt, 1990), it seems clear that there is a need for further improvements. The present results suggest that although DECtalk may accurately model some of the low frequency source characteristics, there may be a problem in the shape of the glottal pulse as revealed by sensitivity to the higher frequency components. Although the present study indicates that variability in glottal pulses is not contributing substantially to the perception of naturalness, we believe that more research is needed to assess this question.

In our second experiment, we found clear evidence that listeners are sensitive to the prosodic structure of spoken words, even when they cannot identify those words. Even for a single syllable, listeners reliably distinguish between human speech and synthetic speech. Again, as in the first experiment, listeners also can reliably distinguish between two types of synthetic speech. Since segmental intelligibility is not a factor in the naturalness judgments (due to low-pass filtering), the perceived differences in naturalness between natural and synthetic speech and between the synthesizers must be due to differences in the prosody of these words. Given that there were no reliable differences as a result of the syllable structure of the words (i.e., monosyllables and polysyllables produced the same results), this suggests that the naturalness differences may be due in part to the pitch contour of the word and its amplitude envelope.

Clearly it is possible to measure reliably the naturalness of synthetic speech. Further, by controlling intelligibility (e.g., through low pass filtering), it is possible to reduce confounding of segmental intelligibility and focus on particular aspects of synthetic speech that reduce the perception of naturalness. These techniques can now be applied to other linguistic units such as phrases, sentences, and passages to focus on other prosodic contributions to naturalness.

Although there have been large improvements in segmental intelligibility of synthetic speech in the past decade, improving the naturalness of speech generated by rule remains a significant challenge. The present experiments represent a substantial first step in developing new standardized diagnostic tests that can measure naturalness. By focusing on particular areas of spoken language, these tests can be used to guide improvements in the naturalness of synthetic speech, thereby increasing the utility of this important interface technology.

Acknowledgments

This research was supported, in part, by a research grant from the Digital Equipment Corporation to the University of Chicago, and in part, by a research grant from the Social Sciences Division of the University of Chicago.

We would like to thank Tony Vitale for numerous discussions about the issues addressed in this paper. Tony has been a continual source of support, encouragement, enthusiasm, and assistance in all phases of this research.

Requests for reprints should be addressed to Howard Nusbaum, Department of Psychology, 5848 South University Avenue, The University of Chicago, Chicago, IL, 60637.

References

Best, C.T., Morongiello, B., and Robson, R. (1981). Perceptual equivalence of acoustic cues in speech and nonspeech perception. *Perception & Psychophysics*, *29*:191–211.

Bollinger, D. (1989). *Intonation and its uses: Melody in grammar and discourse*. Stanford University Press, Stanford.

Campbell, W.N. and Isard, S.D. (1991). Segment durations in a syllable frame. *Journal of Phonetics*, *19*:37–47.

Carrell, T.D. (1984). *Contributions of fundamental frequency, formant spacing, and glottal waveform to talker identification*. Research on Speech Perception Technical Report No. 5, Speech Research Laboratory, Indiana University, Bloomington.

Cooper, W.E. and Paccia-Cooper, J. (1980). *Syntax and speech*. Harvard University Press, Cambridge.

Cooper, W.E. and Sorenson, J.M. (1981). *Fundamental frequency in sentence production*. Springer-Verlag, New York.

Fant, G. (1991). What can basic research contribute to speech synthesis? *Journal of Phonetics*, *19*:75–90.

Flanagan, J.L. (1955). A difference limen for vowel formant frequency. *Journal of the Acoustical Society of America*, *27*:613–617.

Flanagan, J.L. (1957). Difference limen for formant amplitude. *Journal of Speech and Hearing Disorders*, *22*:202–212.

Flege, J.E. (1988). Factors affecting degree of perceived foreign accent in English sentences. *Journal of the Acoustical Society of America*, *84*:70–79.

Holmes, J.N. (1961). Research on speech synthesis. Joint Speech Research Unit Report JU 11-4, British Postal Office, Eastcote, England.

Holmes, J.N. (1973). Influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer, *IEEE Transactions on Audio and Electroacoustics*, *AU-21*:298–305.

House, A.S., Williams, C.E., Hecker, M.H.L., and Kryter, K.D. (1965). Articulation testing methods: Consonantal differentiation with a closed response set. *Journal of the Acoustical Society of America*, *37*:158–166.

Hunnicutt, S. (1995). The development of text-to-speech technology for use in communication aids. In A. Syrdal, R. Bennett, and S. Greenspan (Eds.), *Applied speech technology*, CRC Press, Boca Raton, pp. 547–563.

Klatt, D.H. (1976). Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, *59*:1208–1221.

Klatt, D.H. and Klatt, L.C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America, 87*:820–857.

Kohler, K.J. (1991). Prosody in speech synthesis: The interplay between basic research and TTS application. *Journal of Phonetics, 19*:121–138.

Laver, J. (1980). *The phonetic description of voice quality,* Cambridge University Press, Cambridge.

Lee, L. and Nusbaum, H.C. (1989). The effects of perceptual learning on capacity demands for recognizing synthetic speech. Paper presented at the Acoustical Society of America, Syracuse, May.

Lieberman, P. and Crelin, E.S. (1971). On the speech of Neanderthal man. *Linguistic Inquiry, 2*:203–222.

Lieberman, P., Crelin, E.S., and Klatt, D.H. (1972). Phonetic ability and related anatomy of the newborn, adult human, Neanderthal man, and the chimpanzee. *American Anthropologist, 74*:287–302.

Lisker, L. and Abramson, A.S. (1967). Some effects of context on voice onset time in English stops. *Language & Speech, 10*:1–28.

Logan, J.S., Greene, B.G., and Pisoni, D.B. (1989). Segmental intelligibility of synthetic speech produced by rule. *Journal of the Acoustical Society of America, 86*:566–581.

Nusbaum, H.C., Schwab, E.C., and Pisoni, D.B. (1984). Subjective evaluation of synthetic speech: Measuring preference, naturalness, and acceptability. Research on Speech Perception Progress Report No. 10, Speech Research Laboratory, Department of Psychology, Indiana University, pp. 391–407.

Nusbaum, H.C. and Pisoni, D.B. (1985). Constraints on the perception of synthetic speech generated by rule. *Behavior Research Methods, Instruments & Computers, 17*:235–242.

Pisoni, D.B., Nusbaum, H.C., Luce, P.A., and Slowiaczek, L.M. (1985). Speech perception, word recognition and the structure of the lexicon. *Speech Communication, 4*:75–95.

Pols, L.C.W. and van Bezooijen, R. (1991). Gaining phonetic knowledge whilst improving synthetic speech quality? *Journal of Phonetics, 19*:139–146.

Ralston, J.V., Pisoni, D.B., and Mullennix, J.W. (1995). Perception and comprehension of speech. In A. Syrdal, R. Bennett, and S. Greenspan (Eds.), *Applied speech technology*, CRC Press, Boca Raton, pp. 233–288.

Schmidt-Nielsen, S. (1995). Intelligibility testing and acceptability testing for speech technology. In A. Syrdal, R. Bennett, and S. Greenspan (Eds.), *Applied speech technology*, CRC Press, Boca Raton, pp. 195–232.

Selkirk, E.O. (1986). *Phonology and syntaz: The relation between sound and structure*, MIT Press, Cambridge.

Slowiaczek, L.M. and Nusbaum, H.C. (1985). Effects of speech rate and pitch contour on the perception of synthetic speech. *Human Factors, 27*:701–712.

Schwab, E.C., Nusbaum, H.C., and Pisoni, D.B. (1985). Effects of training on the perception of synthetic speech. *Human Factors, 27*:395–408.

Syrdal, A.K. (1989). Improved duration rules for text-to-speech synthesis. *Journal of the Acoustical Society of America, 85*, S43.

Voiers, W.D. (1977). Diagnostic Acceptability Measure for speech communication systems. IEEE International Conference on Acoustics, Speech, and Signal Processing, New York.

Voiers, W.D. (1983). Evaluating processed speech using the Diagnostic Rhyme Test. *Speech Technology*, pp. 30–39.