APPENDIX 12

ASSESSING THE INTELLIGIBILITY AND QUALITY OF SPEECH

A12.1 INTRODUCTION

The purpose of this appendix is to review factors that can affect the assessment of synthesized speech (Childers and Wu, 1990).

A12.2 INTELLIGIBILITY

The intelligibility of speech may be measured by the ability of listeners to correctly identify aspects of a speech stimulus such as phonemes, syllables, words, or sentences, given that the language is known to the listener and that the syntax and semantics are correct (Rothauser et al., 1968, 1969, 1971). Intelligibility may be rated using an articulation index that counts the number of speech stimulus units correctly identified by a listener group (Fletcher and Steinberg, 1929; French and Steinberg, 1947; Quackenbush et al., 1988). Steeneken and Houtgast (1980) have extended the articulation index by developing the speech transmission index, which uses synthetic signals to test the intelligibility of a communication channel or an auditorium. Several rhyme tests measure the intelligibility of consonants in rhyming words; for example, the rhyme test (Fairbanks, 1958), the modified rhyme test (House et al., 1965; Huggins et al., 1985; Logan et al., 1989), and the diagnostic rhyme test (Voiers, 1968, 1977a, 1977b, 1983). Phonetically balanced word stimuli, rather than rhyming words, may be used to test intelligibility (House et al., 1965; Kryter, 1970). The application of some of these tests to speech coders is discussed by Jayant and Noll (1984), Quackenbush et al. (1988), and more recently by Kitawaki (1992). The latter paper provides an excellent description of the assessment of speech quality for waveform coding and analysis–synthesis. See Hawley (1977) for references to the older literature. The use of such tests for assessing the quality and intelligibility of synthetic speech is discussed later.

A12.3 QUALITY

While there seems to be no universally accepted definition of the quality of speech, it is usually referred to as the total auditory impression the listener experiences upon hearing the speech of another talker. This impression may often be called naturalness, which is a highly subjective attribute conveying the impression "human sounding." The listener's impression is influenced by factors that affect the clarity of the speech such as articulatory proficiency, loudness, vocal quality, resonance, speaking rate, fluency, stress, and intonation. The speaker's age, sex, and dialect affect the listener's impressions as well. Not to be forgotten are environmental factors including interference, such as ambient noise, other signals or competing speech, interfaces such as telephone connections, signal distortions as introduced in helium speech, and so forth. Sometimes researchers use the terms quality, naturalness, and intelligibility as interchangeable. There have been attempts to standardize the

descriptive terminology for the phonetic description of voice quality (Laver, 1980) and to classify voice qualities (Murry and Singh, 1980; Singh and Murry, 1978), but more work is needed in this area, particularly at the interface between speech technology and speech science.

The term quality has been used in different contexts. A phonetician might use quality to describe articulatory differences, as when comparing the vowels in different words. A speech pathologist might speak of laryngeal quality (hoarse, harsh, breathy) or of resonance (vocal tract) and nasal quality. A singer may use quality to express differences in vocal registers, which are related to laryngeal vibratory characteristics. Researchers have attempted to assess speech quality in terms of loudness (Rothauser et al., 1968, 1971). In quality assessment tasks, the intelligibility of individual phonemes may be a major factor listeners attend to (Pisoni et al., 1985). Quality may be determined from a listener's appraisal of a speech stimulus, using comparisons between a test stimulus and a reference utterance of known attributes such as "breathy," "crisp," "rough," and so on. (Colton and Estill, 1981; Logan et al., 1989; Rothauser et al., 1971). A listener's assessment of synthetic speech may be achieved by comparing synthetic speech tokens with tokens of natural speech.

A12.4 ASSESSING SYNTHETIC SPEECH

From a technical point of view, one prefers to have an objective method for assessing intelligibility, naturalness, or quality because the results obtained by such methods are presumably reproducible. Although the acoustic correlates of quality and naturalness are poorly understood, several attempts have been made to design an objective evaluation procedure using factors derived from the speech signal (Barnwell, 1979; Eskenazi, 1988; Eskenasi et al., 1990; Itoh et al., 1984; Jayant and Noll, 1984; Juang et al., 1982; Juang, 1984; Kitawaki, 1992; Koljonen and Karjalainen, 1984; Makhoul et al., 1976; Naik, 1984; Nakatsui and Mermelstein, 1982; Nocerino et al., 1985; Quackenbush et al., 1988; Steeneken and Houtgast, 1980; Viswanathan et al., 1984). A distance metric or distortion measure, based on a set of signal features, is the foundation for these objective measures. Two speech tokens are compared using identical feature sets extracted from each token. The quality is determined by measuring the distance between these feature sets with a distance metric. By selecting a perceptually consistent set of features, one hopes to achieve a high degree of correlation between objective distance measures and listener (or subjective) ratings of the same tokens (Makhoul et al., 1976, 1985; Markel and Gray, 1976). Typical distance measures use spectral data, employing a logarithmic magnitude scale, to assess the difference between two spectra (a reference sample and a test sample). Distance measures have used cepstral coefficients and a log likelihood ratio of spectra. The most frequently used spectral distance measure is perhaps the Itakura–Saito distance (Juang, 1984; Markel and Gray, 1976; Nocerino et al., 1985; Rabiner and Schafer, 1978). Many intelligibility experiments have been concerned with the evaluation of system distortions of natural speech; for example, (1) the effects of signal-to-noise ratios; (2) transmission bandwidth restrictions; (3) clipping and filtering; and (4) other signal distortions (Jayant and Noll, 1984; Licklider and Pollack, 1948; Quackenbush et al., 1988). Such studies, even though they may use distance or distortion measures, usually do not examine intelligibility issues as they relate to speech analysis–synthesis systems. One problem with a distance or distortion measure is that, on occasion, a decrease in the magnitude of the measure of a few decibels may be perceived by the ear quite easily but not so at other times (Makhoul et al., 1985). Consequently, distance measures do not always correlate well with listener evaluations of speech (natural or synthetic).

For synthesized speech the study of intelligibility has typically examined the relative contributions of various analysis and synthesis parameters. Wong and Markel (1977) evaluated the intelligibility of the linear predictive vocoder by systematically varying the parameters of predictor order, frame rate, and frame size. Their evaluation was accomplished using a subjective diagnostic rhyme listening test. They found that voiced speech segments were rated consistently as more intelligible than unvoiced segments. If the predictor coefficients for unvoiced segments were calculated more frequently (e.g., every 11.25 msec instead of every 22.5 msec), then the intelligibility of unvoiced segments improved but was still less than that for voiced segments.

Both the naturalness and intelligibility of synthesized sentences were evaluated by Carlson et al. (1979). They used MIT's text-to-speech system and Klatt's (1980) formant synthesizer. Their study

focused on durational factors in speech synthesis; for example, vowel and consonant duration, and showed that good durational structure for a sentence is important both for naturalness and intelligibility. Other studies have compared the intelligibility of natural speech to the speech of text-to-speech synthesizers (Hoover et al., 1987; Logan et al., 1989; Luce et al., 1983; Pisoni et al., 1985; Pisoni and Hunnicut, 1980). Generally, the error rates for synthetic speech were higher than those obtained with natural speech. One must remember that the intelligibility and quality of speech production by text-to-speech systems is greatly affected by the phonetic and linguistic rules employed by the system.

Context affects the intelligibility of speech (Mack and Gold, 1985; Pisoni, 1982; Pisoni and Hunnicut, 1980). Word recognition using nonsense sentences was less for synthetic speech than for natural speech, presumably because of the information provided by the meaning of the sentence.

Cooper (1987) notes from the work of Nooteboom (1983) and Waterworth (1983) that the judicious placement of pauses at appropriate grammatical positions in sentences may increase the intelligibility of the speech through improved listener recall. This technique seems to work as well for computer generated telephone numbers, as for waveform coders. But one need not manipulate characteristics of sentences of synthetic speech to improve a listener's assessment performance. For example, a listener's evaluation scores will become better as familiarity with (or learning to understand) synthetic speech increases (Schwab et al., 1985).

Manipulating the formants in a synthesizer can affect the intelligibility of the synthesized speech. Suppressing the magnitude of the second formant degraded intelligibility much more so than suppressing the first formant (Agrawal and Lin, 1975). This is in agreement with Thomas (1968), who attributed the high intelligibility of clipped speech to the presence of a strong (relatively unaffected) second formant.

A speaker's vocal characteristics apparently influence the quality of reproduction of the speaker's voice. Using a linear predictive coding (LPC) synthesizer Kahn and Garst (1983) found that male voices were more intelligible than female voices and that the presence of nasality or whisper in the speaker's voice degraded the intelligibility of the synthesized voice for both males and females.

Smith et al. (1981) found that the intelligibility of eigenparameter encoded speech was not improved over LPC speech and that synthetic speech quality was dependent on the eigenparameter quantization technique used. These investigators also found, as did Kahn and Garst (1983), that the LPC synthesizer does not model all voices equally well. They concluded that the LPC model needs to be improved if the synthetic speech is to approach that of natural speech in both intelligibility and quality, a point noted by Atal and David (1979) and Wong (1980).

A12.5 METHODS FOR ASSESSING INTELLIGIBILITY AND QUALITY

From the previous review, it appears that speech naturalness, quality, and intelligibility cannot be characterized as orthogonal coordinates in a three-dimensional space. As we presently use these terms, factors that affect one concept may affect the other two. So a definition of speech quality is needed.

One simple definition of speech quality has been suggested by Rothauser et al. (1968, 1969, 1971). They described speech quality in terms of only four factors: intelligibility, preference, loudness, and speaker recognizability. With the exception of loudness, none of these factors is directly related to a single measurement procedure. We suggest focusing on the factor "preference," a term that describes the average attitude of a listener toward a speech signal while he or she compares it with another speech signal. Preference tests can provide an answer to the question: Which one of two speech tokens to be compared is preferred by an average listener? According to Rothauser et al. (1968, 1969, 1971), the aspect of preference with respect to the overall speech quality becomes dominant when all of the following conditions are fulfilled, which appears to occur in practical situations.

1. The intelligibility of the speech signal is sufficiently high, so that it loses its importance as a prime quantitative speech quality factor and design criterion.
2. The level of the speech signal is maintained at an optimum loudness, eliminating the influence of loudness as a quality factor for the speech signals to be evaluated. Optimum loudness is

specified as the average speech level at which the single listener or a listener group prefers the sound level of the speech presented for the task to be performed.

3. The recognizability of the speaker is of no interest to the listener for the listening task. This is usually the case when the listener is expected to gain no further information from the speech signal than he or she might get from reading the written text.

Under these circumstances, preference may be said to represent speech quality for all practical purposes. Rothauser et al. (1969) suggested that the listener's preference may be expressed as the proportion of the listening group in percent that prefers the speech test signal to the speech reference signal as a source of information. The ratings are typically plotted as a reference scale or expressed as the number of times a signal was preferred or as a percentage of times the signal was preferred. The listener should be capable of discriminating speech quality and be able to express his or her preference in a consistent way. The degree to which the listeners' responses are consistent may be judged by correlational techniques, and used to eliminate listeners whose responses produce highly inconsistent data (Rothauser et al., 1969). Listeners can consider quality as equivalent to naturalness under the conditions discussed in the list presented in this section.

A12.6 LISTENING TESTS

Following Rothauser et al. (1968, 1969, 1971), we suggest that listeners hear two successive speech tokens, which can be sentences or other speech tokens. The listeners can be told the content of the sentences (tokens) that they will hear. The listeners should be given instructions. For example, that they are to indicate their preference for the sentence (token) that sounded the most natural, where naturalness is defined as "human sounding," or a similar preference task.

The speech tokens are usually presented in forced choice pairs. The subjective evaluation will yield a rating based on perceptually defined quantities rather than signal characteristics that can be measured. By presenting the synthesized speech material in a systematic manner, one hopes to identify the synthesis conditions that are "preferred" by a group of listeners. To ensure that the listeners can make the required discriminations in a consistent manner, each listener's performance should be graded. The order of presentation for each pair should be randomized for the complete experiment. For example, the listener might hear the speech token sequence A followed by B. Later, the sequence B followed by A would be presented. For the listener's scoring to be included in a study, a listener is to be consistent at the 75% level or better for all paired comparisons. For example, if only eight presentations of four different paired signals are made (A-B, C-D, E-F, G-H, B-A, D-C, F-E, H-G), then the listener could have at most only one disagreement among the paired signals.

One example of the presentation format for a listening test is a paired comparison in which three synthesized sentences for a particular speaker are compared with each other and with the original sentence, excluding same sentence comparisons. This gives six possible pairs. Each pair is presented twice in each listening session. The pair is also presented twice in reverse order. The order of presentation is randomized. Thus, a total of 24 pairs of sentences are presented for each speaker type. The listeners are told the content of the three sentences that they will hear.

A master listening tape might be generated or a computer stored equivalent. A tone of 500 Hz might be used to cue the listener that the speech material is to follow. The tone is followed by each pair of sentences (A, B), repeated twice; that is, (A, B), (A, B). There is a 1-second interval between the tone and the first A and the subsequent B. This is followed by a 2-second interval before the second A, followed by a 1-second interval before the second B. A 4-second interval is then provided for the listeners to make and score their choice before the next presentation occurrs. No ties in choice are allowed. The results of the listening test are rated usually as a percentage, representing the number of times each stimulus is preferred as more natural sounding when compared with all other stimuli. For example, a rating of 60% means that a particular sentence is preferred as sounding more natural 60% of the time. A rating of 50% would mean that the two sentences being compared would be equally preferable.

Another rating, referred to as the relative preference rating, represents the number of times sentence A is preferred to sentence B, expressed as a percentage of their joint occurrences in both

forward (A, B) and reverse (B, A) order. For example, a relative preference rating of 60% for sentence A compared with sentence B means that out of all the joint occurrences of the two sentences, A is preferred to B 60% of the time.

A12.7 SOME FACTORS THAT AFFECT THE QUALITY OF SYNTHETIC SPEECH

Several factors that affect the quality of synthetic speech produced by analysis–synthesis are: (1) formant locations and bandwidths (or number of poles and their positions); and (2) the excitation waveshape, which may include source–tract interaction. These factors are affected by the analysis techniques employed by the investigator. Furthermore, the manner in which these factors vary dynamically over time must be tracked by the analysis procedure (Childers and Wu, 1990).

Factors that affect LPC speech are modeling errors, inaccuracies in analysis, pitch measurement and modeling errors, errors in voicing detection, and LPC parameter quantization (Childers and Wu, 1990). "Buzziness" in LPC synthesized speech can be due discontinuities between speech segments introduced by segment-by-segment changes that occur in LPC parameters (Childers and Wu, 1990; Kuwabara, 1984). To cure this problem, Kuwabara (1984) suggests smoothing the synthesized speech segments pitch synchronously. The "warbling" in some LPC synthetic speech is often caused by improper measurement and modeling of the gain parameter between speech frames (Childers and Hu, 1994).

Formant synthetic speech can be improved by changing the formant bandwidths with time to account for source–tract interaction and/or by adjusting the source waveform to reflect the affect of formant interaction with the source (Childers, 1995; Childers and Ahn, 1995; Childers and Lee, 1991; Childers and Wong, 1994; Childers et al., 1989; Pinto et al., 1989).

Source excitation characteristics are important for speech synthesis; for example, jitter and shimmer, excitation waveform shape, and source–tract interaction (Childers, 1995; Childers and Ahn, 1995; Childers and Ding, 1991; Childers and Hu, 1994; Childers and Lee, 1991; Childers and Wong, 1994; Childers and Wu, 1990; Childers et al., 1989; Eskenasi et al., 1990; Lalwani and Childers, 1991a and 1991b; Pinto et al., 1989).

A12.8 SUMMARY

The purpose of this appendix is to acquaint the reader with some aspects of the assessment of speech intelligibility and quality and to present one definition for speech quality. In addition, an outline of an example listening test and its evaluation is given. For a discussion of similar factors for speech coding see Kitawaki (1992), which provides a description of the assessment of speech quality for waveform coding and analysis–synthesis. The Kitawaki paper discusses several assessment methods, subjective and objective methods of assessment of coded speech, distortion measures, speaker dependence on quality, and other factors. Several other references related to speech coding include Atal, Cuperman, and Gersho (1991, 1993), Beker and Piper (1985), Papamichalis (1987), Parsons (1986), Furui (1989), Furui and Sondhi (1992), Saito (1992), Singhal and Atal (1989), Tetschner (1993), Witten (1982).