1 Lecture Outline

Reading: Chapter 12

- Speech quality measures

2 Evaluation of Speech Coders

This section is taken from *Speech Enhancement Theory and Practice* by Philipos Loizou.

There are many assessment methods that have been used to evaluate speech coders in terms of speech quality and intelligibility. Quality assessment can be done using subjective listening tests or objective quality measures. *Subjective evaluation* involves comparisons of original and coded/decoded speech signals by a group of listeners who are asked to rate the quality of speech along a predetermined scale, usually from 1 (bad) to 5 (excellent). *Objective evaluation* assesses quality through a "distance measure" between original and decoded speech. Clearly for the objective measure to be valid it needs to correlate well with subjective listening tests.

Quality is only one of many attributes of the speech signal and is highly subjective in nature and is difficult to evaluate reliably. This is partly because individual listeners have different internal standards o what constitutes "good" or "poor" quality resulting in large variability in rating scores among listeners. Quality measures assess "how" a speaker produces an utterance, and includes attributes such as "natural," "raspy," "hoarse," "scratchy," and so on.

Intelligibility is an entirely different attribute and not equivalent to quality. Intelligibility measures "what" the speaker said. Intelligibility is not subjective and can be easily measured by presenting to a group of listeners speech material and asking them to identify the words spoken. Intelligibility is quantified by counting the number of words or phonemes correctly identified.

2.1 Subjective Quality Measures: Mean Opinion Scores

The most widely used direct method of subjective quality evaluation is the category judgement method in which listeners rate the quality of a test signal using a five-point numerical scale ranging from 1 (bad) where the level of distortion is very annoying and objectionable to 5 (excellent) where the level of distortion imperceptible; other scores are 2 (poor), 3 (fair), 4 (good). The measured quality of the test signal is obtained by averaging the scores obtained from all listeners. The average score is referred to as the Mean Opinion Score (MOS).

The MOS test is administered in two phases: training and evaluation (testing). In the training phase, listeners hear a set of reference signals that exemplify the high (excellent), the low (bad), and the middle judgement categories. This phase is very important in subjective evaluation in order to equalize the subjective range of quality ratings of the listeners. In the evaluation phase, subjects listen to the test signal and rate the quality of the signal from 1 to 5.

Although subjective quality measures provide perhaps the most reliable method for assessing speech quality, they can be time-consuming and require (expensive) access to trained listeners.

2.2 Objective Quality Measures

In most objective quality measures, a distance measure is computed between the original speech signal and the processed (encoded/decoded) signal. The distance measure is then usually mapped to a scale from 1 to 5 for comparison to MOS scores. Ideally, the distance measure correlates well with subjective tests.

Objective measures of speech quality are implemented by first segmenting the speech signal into 10-30 ms frames and then computing a distortion measure between original and processed signals. A single, global measure is then computed by averaging the distortion measures of the frames. Distortion measures can be done in time- or frequency-domains.

2.2.1 Segmental SNR Measures

The segmental SNR can be evaluated in the time- or frequency-domain. For this measure to be meaningful, it is important that the original and processed signal be aligned in time (frame-by-frame) and that any phase errors present be corrected. The segmental SNR (SNRseg) is defined as

$$\text{SNRseg} = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\sum_{n=Nm}^{Nm+N-1}x^2[n]}{\sum_{n=Nm}^{Nm+N-1}(x^2[n]-\hat{x}^2[n])} \quad (1)$$

where M is the number of frames, N is the length of the frame, $x[n]$ is the original speech signal, and $\hat{x}[n]$is the processed speech signal. Note that SNRseg is based on the geometric mean of the SNRs across all frames. During intervals of silence, SNRseg may be negative which will bias the overall measure. Therefore, silence frames are usually removed prior to measurement.

As is well known, the human auditory system has a frequency sensitivity which is not uniform, i.e. certain frequencies are perceived louder than others even though they are they same loudness. Fig. **??** illustrates the perceived equal loudness curves.

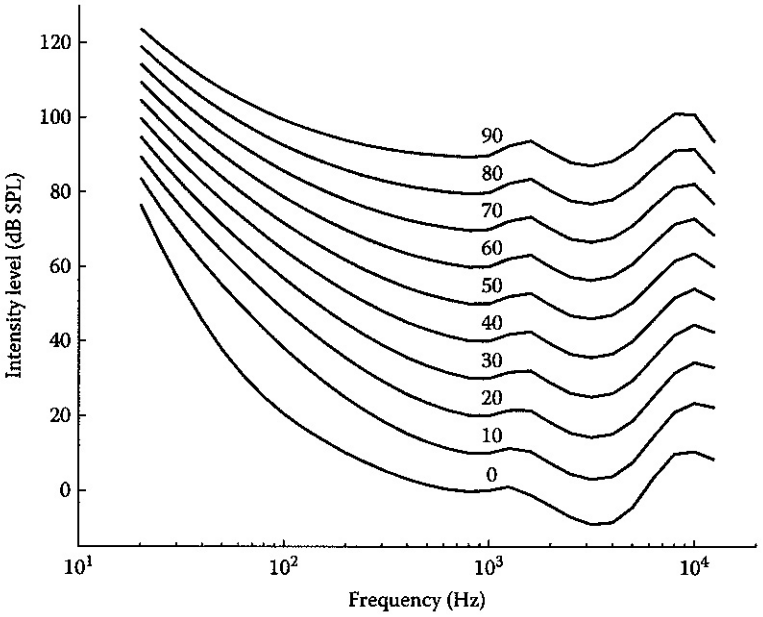

FIGURE 10.6 Equal loudness curves computed using the ISO 226 standard. Numbers on each curve indicate phons.

Figure 1: [Figure 10.6 Loizou] equal loudness curves

Because of the non-uniform frequency sensitivity of the human auditory system, often, the signals are first passed through perceptual weighting filters before computing SNRseg resulting in the perceptually-weighted segmental SNR. See Table 10.21 (Fig. 2) in Loizou for center frequencies and weights. The perceptually-weighted segmental SNR has ben found to yield a high correlation with subjective listening tests.

TABLE 10.21
Center Filter Frequencies (Hz) and Corresponding Articulation Index Weights Used for Computing the Weighted Spectral Distance Measure

Band Number	Center Frequency (Hz)	Weight	Band Number	Center Frequency (Hz)	Weight
1	50	0.003	14	1148	0.032
2	120	0.003	15	1288	0.034
3	190	0.003	16	1442	0.035
4	260	0.007	17	1610	0.037
5	330	0.010	18	1794	0.036
6	400	0.016	19	1993	0.036
7	470	0.016	20	2221	0.033
8	540	0.017	21	2446	0.030
9	617	0.017	22	2701	0.029
10	703	0.022	23	2978	0.027
11	798	0.027	24	3276	0.026
12	904	0.028	25	3597	0.026
13	1020	0.030			

Source: Quackenbush, S., Barnwell, T., and Clements, M. (1988), *Objective Measures of Speech Quality*, Englewood Cliffs, NJ: Prentice Hall, chap. 6.2.

Figure 2: [Table 10.21 Loizou] perceptual weighting filters

2.2.2 Perceptual Evaluation of Speech Quality (PESQ) ITU-T P.862

In 2000 a competition was held by the ITU-T study group to select a new objective measure capable of performing reliably across a variety of codec and network conditions. The Perceptual Evaluation of Speech Quality (PESQ) was selected as the ITU-T recommendation P.862. See Fig. 3 for the block diagram.

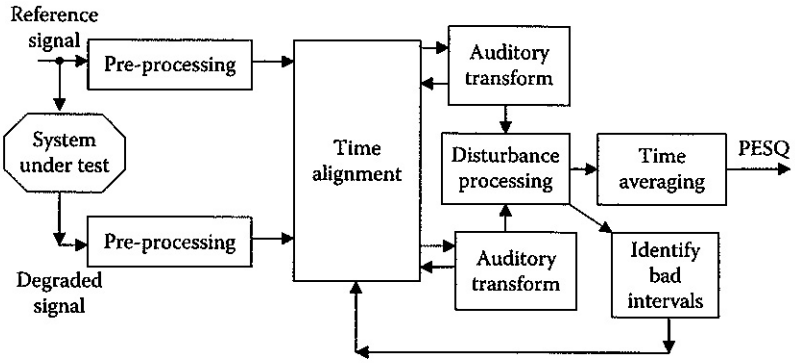

FIGURE 10.9 Block diagram of the PESQ measure computation.

Figure 3: [Figure 10.9 Loizou] block diagram of PESQ algorithm

For the vocoder project, we will use a combination of subjective MOSs and objective PESQ measurements. Students will be provided a code which implements PESQ.