

## Chapter 2

# Speech Quality

There are two aspects to speech quality; the perceived overall speech quality, and the speech intelligibility.

Perceived overall quality is the overall impression of the listener of how “good” the quality of the speech is. The definition of “good” is left to the listener. However, since we hear natural air-transmitted speech emitted from a real humans every day, this speech provides a “reference point” on the quality scale. The listeners rate the speech under test relative to this reference.

On the other hand, speech intelligibility is the accuracy with which we can hear what is being said. The intelligibility is measured as the percentage of the correctly identified responses relative the number of responses. One may use phones, syllables, words or sentences as the test unit. The latter two uses linguistically meaningful units and so care must be taken to use the appropriate and fair choices for the test.

The relationship between perceived quality and speech intelligibility is not entirely understood. However, there does exist some correlation between these two. Generally, speech perceived as “good” quality gives high intelligibility, and vice versa. However, there are samples that are rated as “poor” quality, and yet give high intelligibility scores, and vice versa.

## 2.1 Speech Quality Assessment

Generally, speech quality assessment falls into one of two categories; subjective and objective quality measures. There are subjective and objective measures to measure both of the aspects of speech quality described previously.

Subjective quality measures are based on comparison of original and processed speech data by a listener or a panel of listeners. They rank the quality of the speech according to a predetermined scale subjectively. Evaluation results per listener will include some degree of variation in most cases. This variation can be reduced by

averaging the results from multiple listeners. Thus, results from a reasonable number of speakers need to be averaged for controlled amount of variation in the overall measurement result.

On the other hand, objective speech quality measures are based on some physical measurement, typically acoustic pressure or its electrically converted level in case of speech, and some mathematically calculated values from these measurements. Typically, objective measures are calculated as some distance, typically Euclidean, between objective measurements for the reference speech and the objective measurements for the distorted speech. There are a number of objective measures depending on the application to be tested.

Most of the objective measures have high correlation with subjective measures. Thus, in many cases, one may substitute objective measures to estimate the subjective measures since subjective measurement using listeners is usually much more expensive and time-consuming than objective measurement. However, there are cases where samples with high objective measurement result in poor subjective scores, and vice versa.

## 2.2 Objective Speech Quality Measures

As stated previously, objective speech quality measures are generally calculated from the original undistorted speech and the distorted speech using some mathematical formula. It does not require human listeners, and so is less expensive and less time-consuming. Often, objective measures are used to get a rough estimate of the quality. These estimates are then used iteratively to “screen” subjective quality test conditions so that only the minimum necessary conditions need to be tested subjectively. Many good estimators of subjective quality have been developed, but we still need to evaluate subjective quality at some point since there are still situations where estimations fail.

Some objective quality measures are highly correlated with subjective perceived quality, while others are more correlated with subjective intelligibility. In this section, we will describe a few examples of commonly used objective quality measures.

### 2.2.1 SNR Measures

Signal-to-Noise Ratio (SNR) is one of the oldest and widely used objective measures. It is mathematically simple to calculate, but requires both distorted and undistorted (clean) speech samples. SNR can be calculated as follows:

$$\text{SNR} = 10 \log_{10} \frac{\sum_{n=1}^N x^2(n)}{\sum_{n=1}^N \{x(n) - \hat{x}(n)\}^2} [\text{dB}] \quad (2.1)$$

where  $x(n)$  is the clean speech,  $\hat{x}(n)$  the distorted speech, and  $N$  the number of samples.

This classical definition of SNR is known to be not well related to the speech quality for a wide range of distortions. Thus, several variations to the classical SNR exist which show much higher correlation with subjective quality.

It was observed that classical SNR does not correlate well with speech quality because even though speech is not a stationary signal, SNR averages the ratio over the entire signal. Speech energy fluctuates over time, and so portions where speech energy is large, and noise is relatively inaudible, should not be washed out by other portions where speech energy is small and noise can be heard over speech. Thus, SNR was calculated in short frames, and then averaged. This measure is called the segmental SNR, and can be defined as:

$$\text{SNR}_{\text{seg}} = \frac{10}{M} \sum_{m=0}^{M-1} \log_{10} \frac{\sum_{n=Lm}^{Lm+L-1} x^2(n)}{\sum_{n=Lm}^{Lm+L-1} \{x(n) - \hat{x}(n)\}^2} \quad (2.2)$$

where  $L$  is the frame length (number of samples), and  $M$  the number of frames in the signal ( $N = ML$ ). The frame length is normally set between 15 and 20 ms. Since the logarithm of the ratio is calculated before averaging, the frames with an exceptionally large ratio is somewhat weighed less, while frames with low ratio is weighed somewhat higher. It can be observed that this matches the perceptual quality well, i.e., frames with large speech and no audible noise does not dominate the overall perceptual quality, but the existence of noisy frames stands out and will drive the overall quality lower. However, if the speech sample contains excessive silence, the overall  $\text{SNR}_{\text{seg}}$  values will decrease significantly since silent frames generally show large negative  $\text{SNR}_{\text{seg}}$  values. In this case, silent portions should be excluded from the averaging using speech activity detectors. In the same manner, exclusion of frames with excessively large or small values from averaging generally results in  $\text{SNR}_{\text{seg}}$  values that agree well with the subjective quality. A typical value for the upper and the lower ratio limit is 35 and  $-10$  dB [7]. These ranges are also used for  $\text{SNR}_{\text{seg}}$  calculation throughout this book.

Another variation to the SNR is the frequency-weighted SNR ( $\text{fwSNR}_{\text{seg}}$ ). This is essentially a weighted  $\text{SNR}_{\text{seg}}$  within a frequency band proportional to the critical band. The  $\text{fwSNR}_{\text{seg}}$  can be defined as follows:

**Table 2.1** Weights used in the  $\text{fwSNR}_{\text{seg}}$  calculation

| Band | Center freq. [Hz] | Bandwidth [Hz] | Weights |
|------|-------------------|----------------|---------|
| 1    | 50.000            | 70.0000        | 0.0000  |
| 2    | 120.000           | 70.0000        | 0.0000  |
| 3    | 190.000           | 70.0000        | 0.0092  |
| 4    | 260.000           | 70.0000        | 0.0245  |
| 5    | 330.000           | 70.0000        | 0.0354  |
| 6    | 400.000           | 70.0000        | 0.0398  |
| 7    | 470.000           | 70.0000        | 0.0414  |
| 8    | 540.000           | 77.3724        | 0.0427  |
| 9    | 617.372           | 86.0056        | 0.0447  |
| 10   | 703.378           | 95.3398        | 0.0472  |
| 11   | 798.717           | 105.411        | 0.0473  |
| 12   | 904.128           | 116.256        | 0.0472  |
| 13   | 1020.38           | 127.914        | 0.0476  |
| 14   | 1148.30           | 140.423        | 0.0511  |
| 15   | 1288.72           | 153.823        | 0.0529  |
| 16   | 1442.54           | 168.154        | 0.0551  |
| 17   | 1610.70           | 183.457        | 0.0586  |
| 18   | 1794.16           | 199.776        | 0.0657  |
| 19   | 1993.93           | 217.153        | 0.0711  |
| 20   | 2211.08           | 235.631        | 0.0746  |
| 21   | 2446.71           | 255.255        | 0.0749  |
| 22   | 2701.97           | 276.072        | 0.0717  |
| 23   | 2978.04           | 298.126        | 0.0681  |
| 24   | 3276.17           | 321.465        | 0.0668  |
| 25   | 3597.63           | 346.136        | 0.0653  |

$$\text{fwSNR}_{\text{seg}} = \frac{10}{M} \sum_{m=0}^{M-1} \frac{\sum_{j=0}^{K-1} W(j, m) \log_{10} \frac{X(j, m)^2}{\{X(j, m) - \hat{X}(j, m)\}^2}}{\sum_{j=0}^{K-1} W(j, m)} \quad (2.3)$$

where  $W(j, m)$  is the weight on the  $j$ th subband in the  $m$ th frame,  $K$  is the number of subbands,  $X(j, m)$  is the spectrum magnitude of the  $j$ th subband in the  $m$ th frame, and  $\hat{X}(j, m)$  its distorted spectrum magnitude. An example of the subband allocation and its weight is shown in Table 2.1. These weights were taken from the ANSI SII Standard [3]. There are many variations to the subband definition and the weights. These weights shown in Table 2.1 are also used in the  $\text{fwSNR}_{\text{seg}}$  calculation throughout this book.

Studies have shown that  $\text{fwSNR}_{\text{seg}}$  show significantly higher correlation with subjective quality than the classical SNR or the  $\text{SNR}_{\text{seg}}$  [10, 18].

### 2.2.2 LP-Based Measures

It is well known that the speech production process can be modeled efficiently with a linear prediction (LP) model. There are a number of objective measures that use the distance between two sets of linear prediction coefficients (LPC) calculated on the original and the distorted speech. We will only discuss a few of these.

The Log-Likelihood Ratio (LLR) measure is a distance measure that can be directly calculated from the LPC vector of the clean and distorted speech. LLR measure can be calculated as follows:

$$d_{LLR}(\mathbf{a}_d, \mathbf{a}_c) = \log\left(\frac{\mathbf{a}_d \mathbf{R}_c \mathbf{a}_d^T}{\mathbf{a}_c \mathbf{R}_c \mathbf{a}_c^T}\right) \quad (2.4)$$

where  $\mathbf{a}_c$  is the LPC vector for the clean speech,  $\mathbf{a}_d$  is the LPC vector for the distorted speech,  $\mathbf{a}^T$  is the transpose of  $\mathbf{a}$ , and  $\mathbf{R}_c$  is the auto-correlation matrix for the clean speech.

The Itakura-Saito (IS) distortion measure is also a distance measure calculated from the LPC vector. This measure,  $d_{IS}$  is given by

$$d_{IS}(\mathbf{a}_d, \mathbf{a}_c) = \left[ \frac{\sigma_c^2}{\sigma_d^2} \right] \left[ \frac{\mathbf{a}_d \mathbf{R}_c \mathbf{a}_d^T}{\mathbf{a}_c \mathbf{R}_c \mathbf{a}_c^T} \right] + \log\left(\frac{\sigma_c^2}{\sigma_d^2}\right) - 1 \quad (2.5)$$

where  $\sigma_c^2$  and  $\sigma_d^2$  are the all-pole gains for the clean and degraded speech.

The Cepstrum Distance (CD) is an estimate of the log-spectrum distance between clean and distorted speech. Cepstrum is calculated by taking the logarithm of the spectrum and converting back to the time-domain. By going through this process, we can separate the speech excitation signal (pulse train signals from the glottis) from the convolved vocal tract characteristics. Cepstrum can also be calculated from LPC parameters with a recursion formula.

CD can be calculated as follows:

$$d_{CEP}(c_d, c_c) = \frac{10}{\log 10} \sqrt{2 \sum_{k=1}^P \{c_c(k) - c_d(k)\}^2} \quad (2.6)$$

where  $c_c$  and  $c_d$  are Cepstrum vectors for clean and distorted speech, and  $P$  is the order. Cepstrum distance is also a very efficient computation method of log-spectrum distance. It is more often used in speech recognition to match the input speech frame to the acoustic models.

### 2.2.3 Weighted Spectral Slope Measures

The Weighted Spectral Slope (WSS) distance measure is a direct spectral distance measure. It is based on comparison of smoothed spectra from the clean and distorted speech samples. The smoothed spectra can be obtained from either LP analysis, Cepstrum liftering (a term coined for filtering in the Cepstrum domain), or filter-bank analysis.

One implementation of WSS can be defined as follows,

$$d_{WSS} = \frac{1}{M} \sum_{m=0}^{M-1} \frac{\sum_{j=1}^K W(j, m) (S_c(j, m) - S_d(j, m))^2}{\sum_{j=1}^K W(j, m)} \quad (2.7)$$

where  $K$  is the number of bands,  $M$  is the total number of frames, and  $S_c(j, m)$  and  $S_d(j, m)$  are the spectral slopes (typically the spectral differences between neighboring bands) of the  $j$ th band in the  $m$ th frame for clean and distorted speech, respectively.  $W(j, m)$  are weights, which can be calculated as shown by Klatts in [17].

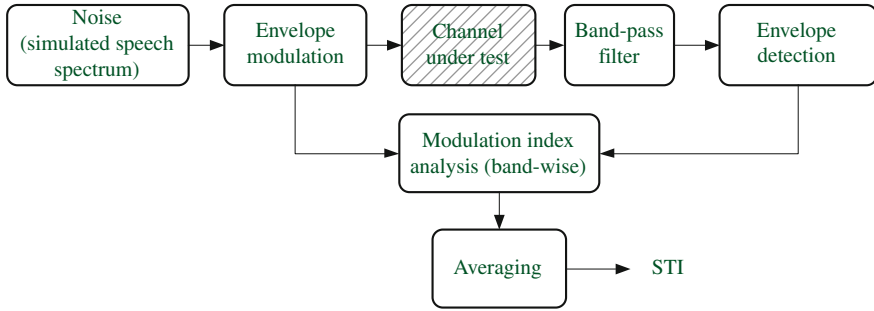
WSS has been studied extensively in recent years, and has enjoyed wide acceptance.

### 2.2.4 Articulation Index

The Articulation Index (AI) was proposed by French and Steinberg [5], and is one of the first widely accepted quality measures that can estimate speech intelligibility. AI assumes that distortions can be calculated on a per-critical frequency band basis, and distortion in one frequency band does not affect other bands. The distortion is assumed to be either additive noise, or signal attenuation. AI can be obtained by calculating the SNR for each band, and averaging them as follows:

$$AI = \frac{1}{20} \sum_{j=1}^{20} \frac{\min\{SNR(j), 30\}}{30} \quad (2.8)$$

where  $SNR(j)$  is the SNR of the  $j$ th subband, the number of subbands is set to 20, and the maximum subband SNR is set to 30 dB. The contribution of each band is set to uniform in this case. The maximum subband SNR can be set to different values, and different weights for each band can be set as well.



**Fig. 2.1** Simplified diagram of the STI measurement

AI can estimate subjective quality well as long as the assumption on the distortion holds. However, many types of distortions are convolutional in nature, and so AI will not be able to estimate quality with these types of distortions well.

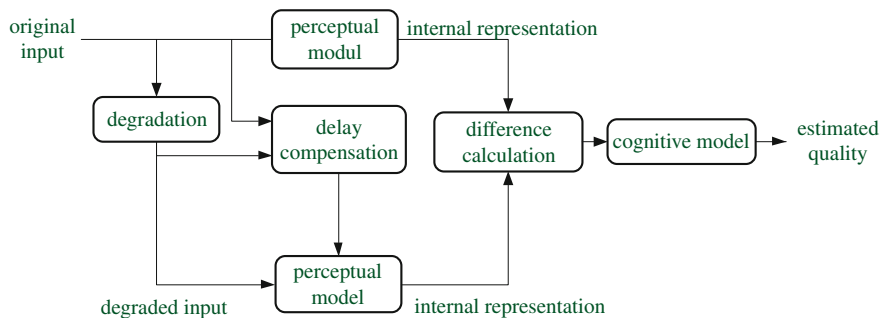
### 2.2.5 Speech Transmission Index

The Speech Transmission Index (STI) is a widely accepted objective measure that can estimate the speech intelligibility for a wide range of environments [20]. Figure 2.1 shows the simplified block diagram of the STI measurement. STI uses an artificial speech signal as input, which is a spectral-shaped noise that has a long-term spectrum envelope identical to speech. This test noise in each band is modulated so that the modulated envelope is sinusoidal. STI assumes that the loss of intelligibility is related to the loss in the modulation depth. The loss in this modulation in each frequency band is calculated, weighed and the averaged at the receiver.

Other objective measures that closely resemble STI exist, such as the Rapid Speech Transmission Index (RASTI), a condensed version of STI, and the Speech Intelligibility Index (SII) [3]. Both of these are known to be equally accurate.

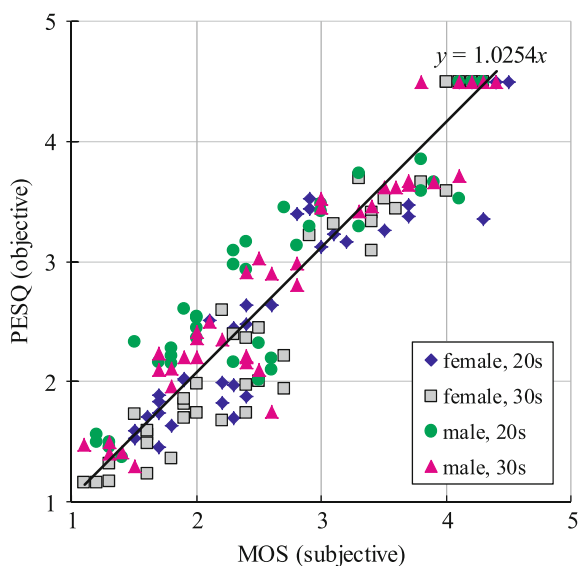
### 2.2.6 PESQ

The Perceptual Evaluation of Speech Quality (PESQ) [13] is an international standard for estimating the Mean Opinion Score (MOS) from both the clean signal and its degraded signal. It evolved from a number of prior attempts to estimate MOS, and is regarded as one of the most sophisticated and accurate estimation methods available today. PESQ was officially standardized by the International Telecommunication Union—Telecommunication Standardization Sector (ITU-T) as standard P.862 in February 2001, and has received some supplementation, including optimized



**Fig. 2.2** Simplified diagram of the PESQ algorithm

**Fig. 2.3** Example MOS estimation using PESQ



mapping to allow more direct comparison with subjective MOS [14], and extension to wideband speech [16]. A simplified diagram of the PESQ is shown in Fig. 2.2.

PESQ uses a perceptual model to convert the input and the degraded speech into an internal representation. The degraded speech is time-aligned with the original signal to compensate for the delay that may be associated with the degradation. The difference in the internal representations of the two signals is then used by the cognitive model to estimate the MOS.

Figure 2.3 is the result of an experiment we conducted to estimate MOS using the PESQ algorithm. We used read Japanese sentences of two male and two female speakers, five per speaker for a total of 20 sentences. White noise was added to these speech samples at 30, 10, and  $-5$  dB. We also encoded and decoded speech samples



with the G.729 CS-ACELP codec [15]. This codec is commonly used in IP telephony applications nowadays. All samples were sampled at 8 kHz, 16 bits per sample. The MOS for all degraded samples were estimated using PESQ. We also ran MOS tests using 10 listeners with the same degraded samples and the original speech. As can be seen in this figure, the estimated MOS generally agrees well with the subjective MOS. The line included in the figure is the fitted line using least mean square error, which came out to be a gradient of 1.024, also showing that the subjective MOS and the estimated MOS generally agrees well.

## 2.3 Subjective Speech Quality Measures

As stated previously in this chapter, subjective quality measures are measures based on the subjective opinion of a panel of listeners on the quality of the speech sample. Generally, subjective quality can be classified into utilitarian and analytical measures. Utilitarian measures results in a measure of speech quality on a uni-dimensional scale, i.e., a numerical value that rate the quality of speech. This numerical value can be used to compare the speech quality resulting from varying conditions, e.g., coding algorithms, noise levels, etc.

On the other hand, analytical measures try to characterize the perceived speech quality on a multidimensional scale, e.g., rough or smooth, bright or muffled. The results of this measure give a value for each of the scale, indicating how the listener perceived the quality on each scale, e.g., how rough or how smooth the listener perceived the test speech sample. In this book, we will only deal with the utilitarian measure.

### 2.3.1 *Opinion Scores*

Opinion rating methods can be used to assess the overall perceived quality of a speech sample. With telephone bandwidth speech, where the bandwidth is limited to between about 300 Hz to 3.4 kHz, the most widely-used opinion rating method is the Mean Opinion Score (MOS) [12]. The listeners rate the speech sample under test into one of the five quality categories, shown in Table 2.2. Each category is assigned a numerical value, also shown in the table. The resulting MOS value is the average value of all listeners for each of the speech under test. Obviously, there are various aspects to the degradation found in the speech under test, e.g., bandwidth limitation, additive noise, echo, nonlinear distortion, etc. MOS results in an overall impression of all these different degradations, measured as one numerical value.

Since the test sample is speech, one can regard the listeners to be using speech they hear from a “live” person as a reference. However, the criteria for each of the quality category are left to the listener. For example, the definition of a “good” speech sample is left to the listener to decide. The weight that each of the listeners assigns on the

**Table 2.2** Speech quality category and five-point scale of the MOS

| Rating | Speech quality category (P.800) | Speech quality (Japanese, translated) | Degradation                       |
|--------|---------------------------------|---------------------------------------|-----------------------------------|
| 5      | Excellent                       | Very Good                             | Imperceptible                     |
| 4      | Good                            | Good                                  | Just perceptible but not annoying |
| 3      | Fair                            | Normal                                | Perceptible, slightly annoying    |
| 2      | Poor                            | Bad                                   | Annoying                          |
| 1      | Unsatisfactory                  | Very Bad                              | Very annoying                     |

various aspects of degradations stated in the previous paragraph will obviously differ. This is why a sizable number of listeners are needed for stable reproducible results. Instructions given to the listeners also can have effect on the results, and so must be carefully controlled. Misleading instruction should not be given. The manner in which the test samples are presented can also have effect on the results, and so should be carefully controlled and maintained to be constant for all listeners. These include the selection of listeners, the ordering of the presented speech under test, type of speech samples in the test set, the presentation method (loudspeakers or headphones, monaural, binaural, diotic, level, etc.), and other environmental conditions.

The quality categorization labels shown in Table 2.2 also are known to have effect on the results. The categorization label in English is standardized in the ITU recommendation P.800 [12]. The categorization label commonly used in Japanese (its direct translation) is also shown in this table. As shown, the labels are not exactly same, and cultural differences give different impressions on the perceived quality. Thus, MOS rating in different languages are known to show some differences even under the same conditions [6].

### 2.3.2 *Speech Intelligibility*

Speech intelligibility tries to measure the accuracy with which the speech under test carries its spoken content. This accuracy depends on the speaker characteristics, the listener, and numerous types of degradation encountered during transmission. It has been used widely to evaluate building or room acoustics, hearing aid performance, speech codec degradation, speech synthesis performance, and many others.

Japanese intelligibility tests often used stimuli of randomly selected single mora, two morae, or three morae syllables [11]. A brief description of these tests will be given in the following section. The subjects were free to choose from any combination of valid Japanese syllables. One can easily see that this will quickly become a strenuous task as the transmission channel distortion increases to an extreme level. Thus, intelligibility tests of this kind is unstable, and often do not reflect the physically evident distortion, giving surprising results [19]. There have been intelligibility

**Table 2.3** An Example syllable table used in the Japanese syllable intelligibility test

| No. | 1   | 2   | 3   | 4   | 5   |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 1   | re  | pa  | ro  | pya | bya | kyo | o   | do  | mi  | ryu |
| 2   | kya | te  | go  | nya | ra  | gya | ru  | a   | pu  | kyu |
| 3   | pyu | me  | ri  | sha | ga  | chi | pyo | ma  | sa  | nyu |
| 4   | hu  | byu | hi  | hyo | ze  | ji  | su  | myo | se  | da  |
| 5   | e   | gyu | gu  | mya | ge  | ya  | bi  | byo | chi | jo  |
| 6   | nyo | zu  | ku  | ho  | cha | mu  | mo  | rya | gi  | ka  |
| 7   | ni  | gyo | bu  | bo  | pe  | cho | zo  | ke  | i   | hya |
| 8   | ja  | u   | ryo | he  | chu | ko  | tu  | so  | ju  | ba  |
| 9   | myu | ta  | sho | ha  | za  | pi  | de  | no  | si  | be  |
| 10  | nu  | wa  | yu  | ne  | ki  | po  | shu | na  | hyu | yo  |

tests that use Japanese word speech as its stimuli [1], but these were generally not widely used.

From early days, English testing of intelligibility has used rhyming words, and listener response was constrained to these rhyming words. The Fairbanks test [4] uses a single syllable rhyming words of the form consonant-vowel-consonant. The listener listens to the valid word speech, and is given a response sheet with the first consonant blank, which they must fill. This testing was further modified to constrain the test material to rhyming word list from which the listener chooses. Details will be described in Sect. 2.3.2.3.

### 2.3.2.1 Syllable Intelligibility

The syllabic intelligibility test uses a random single mora, two morae, or three morae speech to test its listening accuracy. A mora in Japanese is “roughly” equivalent to a syllable. Table 2.3 lists all 100 Japanese morae excluding the syllabic nasal (/N/) and the double (geminate) consonant. This list is randomized, and the listener picks out the correct mora from this table. One can easily see that this quickly becomes a strenuous task as the distortion increases. Thus, intelligibility tests of this kind are known to be unstable, and often give surprising results [11]. Thus, a well-trained listener panel is generally required for stable reproducible result.

### 2.3.2.2 Word and Sentence Intelligibility

Word and sentence intelligibility tests use valid words or sentences as its test material. Word intelligibility is measured by the correct number of words identified by the listener. Sentence intelligibility tests uses question or command sentences, and is measured by the number of correct responses made by the listener. Sentence intelligibility may also be measured by the correct identification of key words embedded in the test sentences, although there are arguments that this is merely a word intelligi-

bility test. Some sentence intelligibility tests use nonsense (meaningless) sentences to avoid the effect of context, and measure the intelligibility by the identification of nouns embedded in the sentences. Sentence intelligibility tests are known to be time-consuming, and listener learning also needs to be considered. The test material also needs careful preparation. However, sentence intelligibility tests can potentially measure intelligibility that closely matches the actual listening condition.

### 2.3.2.3 Forced-Selection Tests

As described before, English intelligibility tests used rhyming words as its constraint. The Fairbanks test allowed user to fill in any valid initial consonant in its test. House et al. further constrained the material to six rhyming words [8, 9]. The listeners now had to choose one word from a list of six rhyming words. This greatly simplifies the listening task, as well as its administration and scoring. Fifty sets of six-word list were defined for this test, some with different initial consonant, and some with different final consonant. This test is called the Modified Rhyme test (MRT). Voiers further constrained the test material to word-pairs, with different initial consonant [21, 22]. The initial consonants were arranged so that the consonant in one of the word-pair would have, and the other would not have one of the six phonetic feature defined in this test. This test is called the Diagnostic Rhyme Test (DRT), and is now an ANSI standard [2]. The MRT and the DRT will be described in detail in the next section.

Japanese intelligibility tests have traditionally not used rhyming words as its test material. The benefits of using constrained rhyming words as its test material were evident in English tests. Thus, the author defined an intelligibility test using Japanese rhyming words in order to utilize this benefit in Japanese intelligibility testing as well. This test is the main topic of this book.

## 2.4 Conclusion

This chapter briefly described two aspects of speech quality; opinion scores that measure the overall perceived speech quality, and speech intelligibility that measures the accuracy of the received speech content. These two different aspects were described, and two types of measures for these aspects; the subjective and objective quality measures. Subjective quality measures employ human listeners to rate the quality of the speech. It often requires considerable number of listeners to obtain stable results, and is often time-consuming and expensive. On the other hand, Objective quality measures estimate the speech quality from some form of physical measurements. Many are based on distance measures between the original and degraded speech. Signal-to-noise ratios (SNR) or some extension of this measure is a common form of a subjective measure. Recent subjective measures use a more sophisticated distance measures that are known to be highly correlated with human auditory perception.

Subjective measures do not require human listeners, and so is less expensive and less time-consuming than subjective measures. However, there are examples where quality estimations from objective measures and subjective measures do not match. Thus, subjective quality measures are still the most conclusive way to measure the perceived quality. However, recent objective measures are good estimators of subjective quality, and can be used to get a rough quality estimate, which can be followed by subjective quality assessment for selected conditions to confirm the perceived quality.

## References

1. Akabane, M., Itahashi, S.: Performance evaluation methods for speech synthesis systems. In: Proceedings of the Acoustical Society of Japan Fall Convention, pp. 215–218 (2000) (in Japanese)
2. American National Standards Institute (ANSI): Method for measuring the intelligibility of speech over communication systems (ANSI S3.2-1989) (1989)
3. American National Standards Institute (ANSI): Methods for calculation of the speech intelligibility index (ANSI S3.5-1997) (1997)
4. Fairbanks, G.: Test of phonemic differentiation: the rhyme test. *J. Acoust. Soc. Am.* **30**, 596–600 (1958)
5. French, N.R., Steinberg, J.C.: Factors governing the intelligibility of speech sounds. *J. Acoust. Soc. Am.* **19**(1), 90–119 (1947)
6. Goodman, D., Nash, R.D.: Subjective quality of the same speech transmission conditions in seven different countries. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 7, pp. 984–987. Paris, France (1982)
7. Hansen, J.H.L., Pellom, B.L.: An effective quality evaluation protocol for speech enhancement algorithms. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP), vol. 7, pp. 2819–2822 (1998)
8. House, A.S., Williams, C.E., Hecker, M., Kryter, K.D.: Psychoacoustic speech tests: a modified rhyme test. Technical Documentary Report US Air Force System Command (ESD-TDR-63-403), pp. 1–44 (1963)
9. House, A.S., Williams, C.E., Hecker, M., Kryter, K.D.: Articulation-testing methods: consonantal differentiation with a closed-response set. *J. Acoust. Soc. Am.* **37**(1), 158–166 (1965)
10. Hu, Y., Loizou, P.C.: Evaluation of objective quality measures for speech enhancement. *IEEE Trans. Audio Speech Lang. Process.* **16**(1), 229–238 (2008)
11. Iida, S.: On the articulation test. *J. Acoust. Soc. Jpn* **43**(7), 532–536 (1987) (in Japanese)
12. ITU-T: ITU-T Recommendation P.800: Method for Subjective Determination of Transmission Quality (1996)
13. ITU-T: ITU-T Recommendation P.862: Perceptual Evaluation of Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs (2001)
14. ITU-T: ITU-T Recommendation P.862.1: Mapping Functions for Transforming P.862 Raw Result Scores to MOS-LQO (2003)
15. ITU-T: ITU-T Recommendation G.729: Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear Prediction (CS-ACELP) (2007)
16. ITU-T: ITU-T Recommendation P.862.2: Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs (2007)
17. Klatt, D.: Prediction of perceived phonetic distances from critical band spectra. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 7, pp. 1278–1281. Paris, France (1982)

18. Ma, J., Hu, Y., Loizou, P.C.: Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *J. Acoust. Soc. Am.* **125**(5), 3387–3405 (2009)
19. Nishimura, R., Asano, F., Suzuki, Y., Sone, T.: Speech enhancement using spectral subtraction with wavelet transform. *IEICE Trans. Fundam.* **79-A**(12), 1986–1993 (1996) (in Japanese)
20. Steeneken, H.J.M., Houtgast, T.: A physical method for measuring speech transmission quality. *J. Acoust. Soc. Am.* **67**(1), 318–326 (1980)
21. Voiers, W.D.: Diagnostic evaluation of speech intelligibility. In: Hawley, M.E. (ed.) *Speech Intelligibility and Speaker Recognition*, pp. 374–387. Dowden, Hutchinson & Ross, Stroudsburg (1977)
22. Voiers, W.D.: Evaluating processed speech using the diagnostic rhyme test. *Speech Technol.* **1**, 30–39 (1983)

Subjective Quality Measurement of Speech  
Its Evaluation, Estimation and Applications

Kondo, K.

2012, XIV, 154 p., Hardcover

ISBN: 978-3-642-27505-0