8

Estimation of Speech Intelligibility Using Perceptual Speech Quality Scores

Kazuhiro Kondo
Graduate School of Science and Engineering, Yamagata University
Japan

1. Introduction

Recent advances in mobile wireless communication devices have made possible speech communication in a variety of noise environments which were not possible before. Also, sophisticated speech encoders, echo control devices, and noise canceling devices have caused artificial synthetic noise, *e.g.* musical noise, which were not seen before with analog or simple PCM speech communication. Thus, a need for comprehensive speech communication quality measures and frequent evaluation efforts have become a necessity. Speech quality is generally measured in one of two measures. The overall listening quality, such as the "naturalness" of the test speech, is typically measured as the Mean Opinion Score (MOS) (ITU-T, 1996). The other criteria is speech intelligibility, which tries to measure the accuracy with which the test speech material carries its spoken content. We will deal mainly with the latter measure in this chapter.

There were not many variations in the types of degradations seen in conventional speech communication systems. Common types of degradations seen were simple ones, such as band limitation and additive noise. Thus, evaluation procedures were fairly simple. Traditionally, Japanese intelligibility tests often used stimuli of randomly selected single mora, two morae or three morae speech (Iida, 1987). The subjects were free to choose from any combination of valid Japanese syllables. This quickly became a strenuous task as the channel distortion increases. Thus, intelligibility tests of this kind is known to be unstable and often do not reflect the physically evident distortion, giving surprising results (Nishimura et al., 1996).

English intelligibility tests are also reported to show similar trends. Accordingly, the Diagnostic Rhyme Test (DRT) (Voiers, 1977; 1983), a closed set selection test that restricted the reply to two words, was proposed. This test is said to be effective in controlling various factors including the amount of training and phonetic context, and is known to give stable intelligibility scores. The DRT has now become an ANSI standard (ANSI, 1989).

In this chapter, we will briefly describe a DRT-type closed set selection test in Japanese (Kondo et al., 2007; 2001). We categorized Japanese consonants into the same taxonomy used for the English tests, and proposed a minimum-pair list accordingly which differ only by the initial consonant and by a single phonetic feature. Subjective test results are also shown with various noise under various SNR.

Then, we will investigate on methods to estimate intelligibility through objective measures. If this is possible with reasonable accuracy, we should be able to "screen" the intelligibility in many of the conditions, and limit the need for full-scale subjective test to a minimum subset.

Feature	m	n	z	ǰ	b	d	g	w	r	j	ɸ	s	š	č	p	t	k	h	N	ts	ç
Voicing (vocalic-nonvocalic)	+	+	+	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	+	−	−
Nasality (nasal-oral)	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−
Sustention (continuant-interrupted)	−	−	+	−	−	−	−	+	+	+	+	+	+	−	−	−	−	+	−	−	+
Sibilation (strident-mellow)	−	−	+	+	−	−	−	−	−	−	−	+	+	+	−	−	−	−	−	+	−
Graveness (grave-acute)	+	−	−	0	+	−	0	+	−	0	+	−	0	0	+	−	0	0	0	−	0
Compactness (compact-diffuse)	−	−	−	+	−	−	+	−	−	+	−	−	+	+	−	−	+	+	−	−	+
Vowel-like (glide-nonglide)	−	−	−	−	−	−	−	+	+	+	−	−	−	−	−	−	−	−	−	−	−

Table 1. The Japanese consonant taxonomy

We will describe our efforts using PESQ (Perceptual Evaluation of Subjective Quality) scores, an ITU standard which estimates MOS from both degraded and original speech, and try to map PESQ-derived MOS to intelligibility.

2. The Japanese diagnostic rhyme test

2.1 Diagnostic rhyme test

Diagnostic Rhyme Tests (DRT) are speech intelligibility tests that forces the tester to choose one word that they perceived from a list of two rhyming words. The two rhyming words differ by only the initial consonant by a single distinctive feature.

DRT assumes the following simplification and principles which will enable even naive listeners to provide stable and efficient intelligibility scores (Voiers, 1977; 1983).

- Additive and convolutional noise mostly affect consonants, which carry the bulk of linguistic information, and not vowels. Thus, exact reproduction of consonants are essential in voice communications. This is also the basis for the Fairbanks Rhyme Tests (Fairbanks, 1958), which tested only consonant recognizability.
- Consonant apprehensibility in the initial, intervocalic and final positions are strongly correlated. Thus, one can measure apprehensibility in all positions just by measuring at the initial position. This assumption is backed by experiments by Suzuki *et al.* (Suzuki et al., 1998), in which they found that there is a strong correlation in the articulation scores of the first and second mora.
- The effect of word familiarity and phonetic context can be neglected if the number of response choices (Miller et al., 1951; Voiers, 1977). In the case of the DRT, the response is restricted to one word out of a pair of words.

In accordance with these assumptions, the DRT uses word-pairs which are minimal pairs in which only the initial consonant differs by a single phonetic attribute as defined by Jakobson, Fant, and Halle (Jakobson et al., 1952). The choice of word-pairs from which the listener selects their response always contains the correct word.

2.2 The Japanese consonant taxonomy

We first proposed a consonant taxonomy for Japanese with the same feature classification used in English, which were drawn from the classification by Jakobson, Fant and Halle (Jakobson et al., 1952) (to be denoted as JFH classification). Table 1 shows the proposed Japanese consonant taxonomy. The "+" shows that the feature is present, the "-" shows the absence,

and "0" shows that the feature does not apply to the consonant. The following seven features were used.

1. Voicing: corresponds to the vocalic-nonvocalic classification by JFH. This is a trivial classification.
2. Nasality: corresponds to the nasal-oral classification by JFH. This is also a fairly trivial classification.
3. Sustention: corresponds to the continuant-interrupted classification. This classifies consonants into clearly continuous consonants and other transient phones, such as plosives.
4. Sibilation: corresponds to the strident-mellow classification. This roughly corresponds to the randomness of the consonants.
5. Graveness: corresponds to the grave-acute opposition. If the spectrum of the consonant concentrates in the low frequency region, it is classified as grave, and vice versa. Also, the oral cavity is not obstructed with grave consonants, while with acute consonants, the oral cavity is divided into compartments with the tongue.
6. Compactness: corresponds to the compact-diffuse opposition. If the spectrum of the consonant largely concentrates around the formant, it is classified as compact, and vice versa.
7. Vowel-like: this classification is not used. It classifies consonants into glides and other true consonants.

We classified most consonants in Japanese speech roughly in the same manner as English. However, several exceptions were noted.

- The consonant [g] is often nasalized in intervocalic positions. However, since we are only dealing with initial consonants, this consonant was classified as oral. Thus, nasality was classified as "-" (feature absent).
- Allophones such as [ŋ] were not classified.

2.3 The Japanese DRT word-pair list

The consonant taxonomy was then used to compile a word-pair list to be used as stimuli for the DRT. Ten word-pairs per each of the 6 features, one pair per each of the five vowel context, were proposed for a total of 120 words (Fujimori et al., 2006; Kondo et al., 2007). The word-pairs are rhyme words, differing only in the initial phoneme. The proposed word-pair list is shown in Table 2. The first words in the word-pair list are words whose initial consonants have the consonant feature under test, and the initial consonants in the latter words do not. Note that all five vowel context are covered.
The following is specific for the Japanese list:

- Only two morae words were initially considered. Longer words will be considered as needed.
- Foreign words were avoided when possible. However, words starting with the [p] context are mostly foreign words, and thus foreign words were included in this case.
- Only words with the same accent type were selected as a word-pair.
- We tried to select mostly common nouns. Proper nouns, slang words and obscure words were avoided where possible.

Voicing	Nasality	Sustention	Sibilation	Graveness	Compactness
Zai - Sai	Man - Ban	Hashi - Kashi	Jamu - Gamu	Waku - Raku	Yaku - Waku
Daku - Taku	Nai - Dai	Hata - Kata	Chaku - Kaku	Pai - Tai	Kai - Pai
Giji - Kiji	Misu - Bisu	Shiri - Chiri	Shiki - Hiki	Mie - Nie	Gin - Bin
Gin - Kin	Miru - Biru	Hiru - Kiru	Chiji - Kiji	Misu - Nisu	Kiza - Piza
Zui - Sui	Muri - Buri	Suki - Tsuki	Chuu - Kuu	Muku - Nuku	Kuro - Puro
Guu - Kuu	Mushi - Bushi	Suna - Tsuna	Jun - Gun	Mushi - Nushi	Yuu - Ruu
Zei - Sei	Men - Ben	Hen - Ken	Shea - Hea	Men - Nen	Gen - Ben
Deba - Teba	Neru - Deru	Heri - Keri	Sheru - Heru	Pen - Ten	Ken - Pen
Zoo - Soo	Mon - Bon	Hoshi - Koshi	Joo - Goo	Moo - Noo	Goki - Boki
Goji - Koji	Nora - Dora	Horu - Koru	Shoji - Hoji	Poro - Toro	Yoka - Roka

Table 2. Japanese DRT word-pair list

- Words which include double consonants and palatalized syllables were excluded when possible.

Additionally, rare consonant-vowel combinations were substituted with other syllables where possible.
As stated before, familiarity may affect the intelligibility scores, although using word-pairs will most likely mitigate this effect. However, to be safe, we selected words which have relatively high phonetic-text familiarity (average 5.5, standard deviation 0.72 on a 7-point scale) according to the familiarity listing compiled by Amano *et al.* (Amano & Kondo, 1999).
Word accents types were judged with reference to both (Amano & Kondo, 1999) and (NHK Broadcasting Culture Research Institute, 1998). Over 77 % of the words were accent type 1 (high to low pitch accent transition), and 2 % were type 0 (flat). Both words in the word-pair had the same accent type. When multiple accent types exist, the speakers were asked to record using the specified accent type, with the same accent type as the other word in the word-pair. The recorded speech was checked for clear pronunciation and accent, and re-recorded as needed.

2.4 The DRT evaluation procedure

Words spoken by multiple speakers should be used. At least 8 listeners should be employed for the test. The listener listens to the stimulus word speech, and selects the correct answer from one of the words in the word-pair. The ordering of the stimulus can be completely random, or it can cycle through the vowel context (*i.e.* form a 5-word cycle covering the five vowel context). The intelligibility is measured by the average correct response rate over each of the six consonant features, or by the average over all features. The correct response rate (CACR) should be calculated using the following formula to compensate for the chance level,

$$S = \frac{100(R - W)}{T} [\%] \tag{1}$$

where S is the response rate adjusted for chance ("true" correct response rate), R is the observed number of correct responses, W the observed number of incorrect responses, and T the total number of responses. In other words, since this is a two-to-one selection test, a completely random response will result in half of the responses to be correct. With the above formula, completely random response will give average response rate of 0 %.

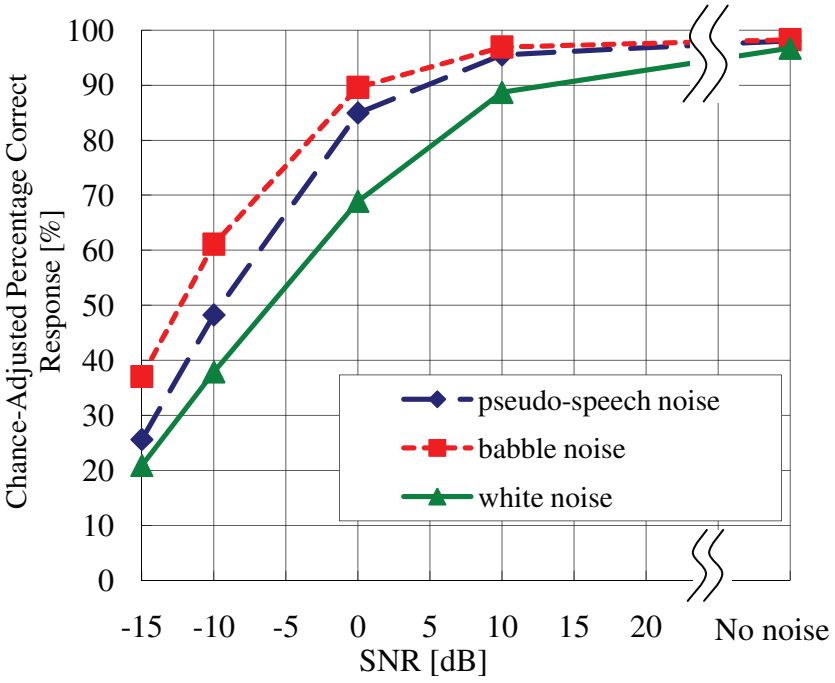

Fig. 1. Comparison of DRT scores for speech with three types of noises

2.5 DRT evaluation experimental setup

We evaluated the DRT on a relatively large Japanese speech database with three typical noise types in order to compare its sensitivity to noise with DRT results in English. We collected speech from eight untrained speakers, four male (all in their twenties) and four female (three in their twenties, and one in her fifties). All words in the DRT word list were recorded using a head-mount electret microphone (Sennheiser HD 410-6) at a sampling rate of 16 kHz, 16 bits per sample. No directions on the pronunciation and accents were initially given, so that the speakers would be able to speak naturally. Re-recordings were made as needed when the speech samples were not of standard accent, or unclear. White noise, multi-speaker (babble) (Rice University, 1995) and pseudo-speech noise (Tanaka, 1989) were mixed into these samples at an SNR of −15, −10, 0 and 10 dB, respectively. Speech for words in the word-pair list was played out in random order. All speech were played out diotically through headphones (Sennheiser HD 25-1 II) at the listener's preferred output level. The listeners were shown both words in the word-pair to choose from. Eleven listeners underwent the tests for speech mixed with white noise, and 5 listeners tested speech in pseudo-speech and babble noise. All listeners were native Japanese speakers in their twenties with reportedly normal hearing. Each listener listened to 8 speakers, 5 noise levels including clean, 6 phonetic features, and 20 words per feature, bringing the total to 4800 spoken words per noise type.

2.6 Results and discussion

Figure 1 shows the average DRT scores (the chance adjusted correct response percentage, CACR) over all phonetic features for the three types of noise tested. Figures 2, 3, and 4 show

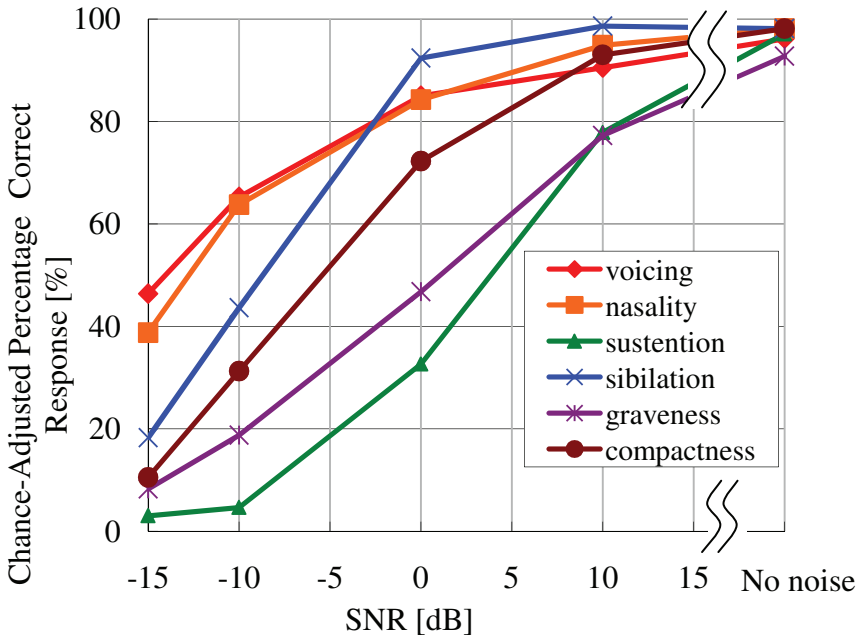

Fig. 2. DRT scores for speech mixed with white noise

CACRs for each of the mixed noise types by the phonetic feature. Two-way ANOVA tests have confirmed SNR and phonetic feature to be main effects in all noise types tested.
The overall trend for all noise types generally agrees with English results obtained by Voiers (Voiers, 1977; 1983). The following can be drawn from the results:

1. The average DRT score over all phonetic feature vs. SNR is similar regardless of the noise type. However, white noise seems to affect the scores most, followed by pseudo-speech noise, and babble. The reason for this seems to be the bandwidth of the noise, especially in the high frequency regions.
2. Sibilation generally shows high scores when white noise level is low. However, the scores decrease quickly as white noise level increases. This again agrees well with results by Voiers (Voiers, 1977; 1983). The reason for this can be that phones with sibilation show wide frequency bandwidth, similar to white noise. This may also be the reason the scores are not affected as much by other types of noise since these have much narrower bandwidth.
3. Much less difference by features is seen with pseudo-speech and babble noise compared to white noise. In other words, each of the phonetic feature is affected similarly with these noise types. Nasality, sustention, and compactness especially show insignificant differences. This was observed in English tests as well. The reason for this again may be the bandwidth of the added noise.

Figure 5 compares the DRT scores for white noise-added speech by speaker gender. As shown by this figure, the DRT scores are virtually same for both male and female speech for all ranges of SNR tested, and thus the gender of the speaker has insignificant effect on the DRT scores. This was also confirmed with ANOVA.

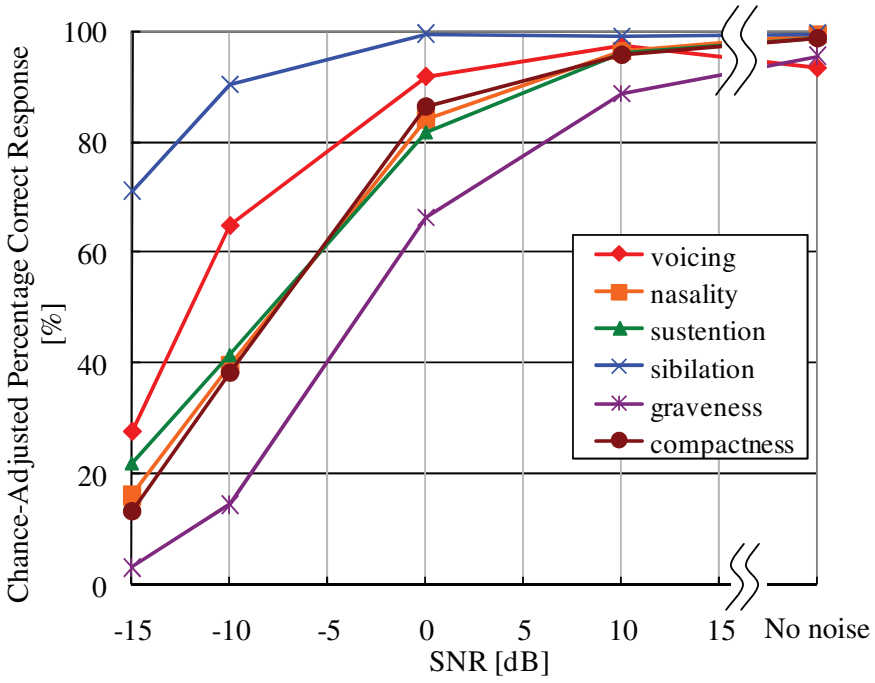

Fig. 3. DRT scores for speech mixed with pseudo-speech noise

3. Estimation of DRT scores using objective measures

In this section, we will describe our approach to estimating the subjective intelligibility DRT scores using objective measures. Even though the proposed DRT tests were much simpler than conventional intelligibility tests, the DRT test still requires human listeners to rate more than one hundred words per noise condition. Accordingly, in the following, we attempted to estimate subjective DRT scores using objective measures obtained by some calculations without human participants. If estimation of intelligibility, at least to some degree, is possible, we should be able to "screen" the intelligibility in many of the conditions, and limit the need for full-scale subjective tests to a minimum.

3.1 Estimation of DRT using PESQ

In this section, we will describe results of experiments to estimate DRT scores from PESQ (Perceptual Evaluation of Speech Quality) scores (Kaga et al., 2006). PESQ is an international standard which tries to estimate subjective Mean Opinion Scores (MOS) (ITU-T, 1996) from the original and the degraded signal (Beerends et al., 2002; ITU-T, 2001; Rix et al., 2002). PESQ is known to be one of the most accurate objective methods to estimate subjective MOS. Although MOS is a subjective measure of the overall speech quality, we can assume that speech quality is "loosely" correlated with speech intelligibility. Thus, we can assume that speech intelligibility is related to estimated MOS values, at least to some degree.

Kitawaki and Yamada have recently conducted a small scale test to employ PESQ to estimate word intelligibility (Kitawaki & Yamada, 2007). They used speech categorized into four classes of word familiarity. They found relatively high correlation between subjective word

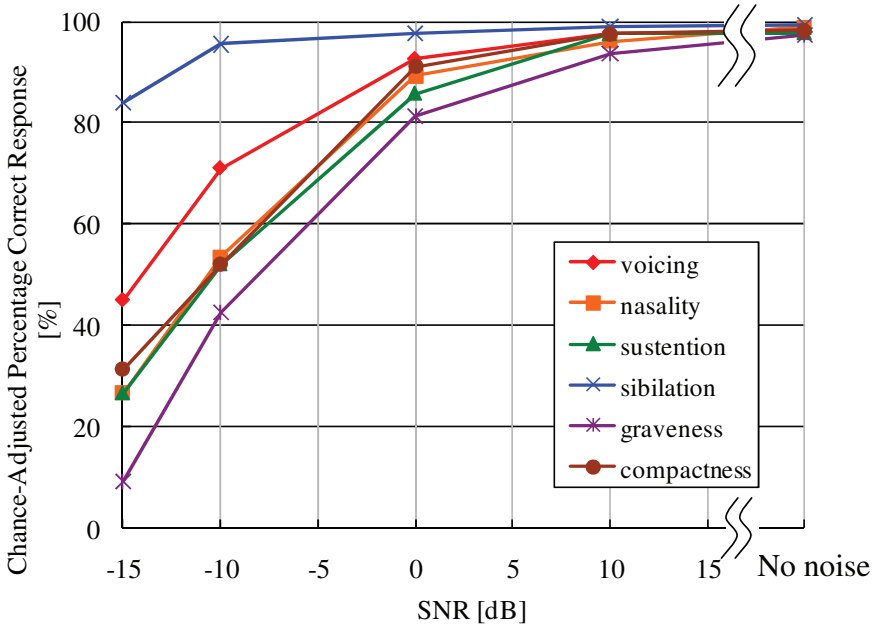

Fig. 4. DRT scores for speech mixed with babble noise

intelligibility and estimated word intelligibility using PESQ scores, especially when the word familiarity is low. Beerends *et al.* also used PESQ to estimate intelligibility (Beerends et al., 2009). They found that PESQ fails to predict intelligibility especially at lower SNR. Thus, they use several methods to improve the estimation in this region, *e.g*, the use of spectral subtraction, silent interval deletion, and steady-state suppression. They show some success in improving the accuracy. On the other hand, Liu *et al.* have recently attempted to estimate speech intelligibility from a number objective measures including PESQ scores (Liu et al., 2008). They used digits for their speech samples, and found very low correlation between intelligibility and PESQ scores. In fact, they found low correlation in most of the objective measures they attempted, highlighting the difficulty of this problem.

3.2 Perceptual Evaluation of Speech Quality (PESQ)

The Perceptual Evaluation of Speech Quality (PESQ) (ITU-T, 2001; 2003; 2005) is an international standard for estimating the Mean Opinion Score (MOS) from both the clean and 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 recommendation P.862 in February, 2001, and extended to wideband speech as recommendation P.862.2 in November, 2005. A simplified diagram of the PESQ is shown in Fig. 6.

PESQ uses a perceptual model to covert 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 two internal representation is then used by the cognitive model to estimate the MOS.

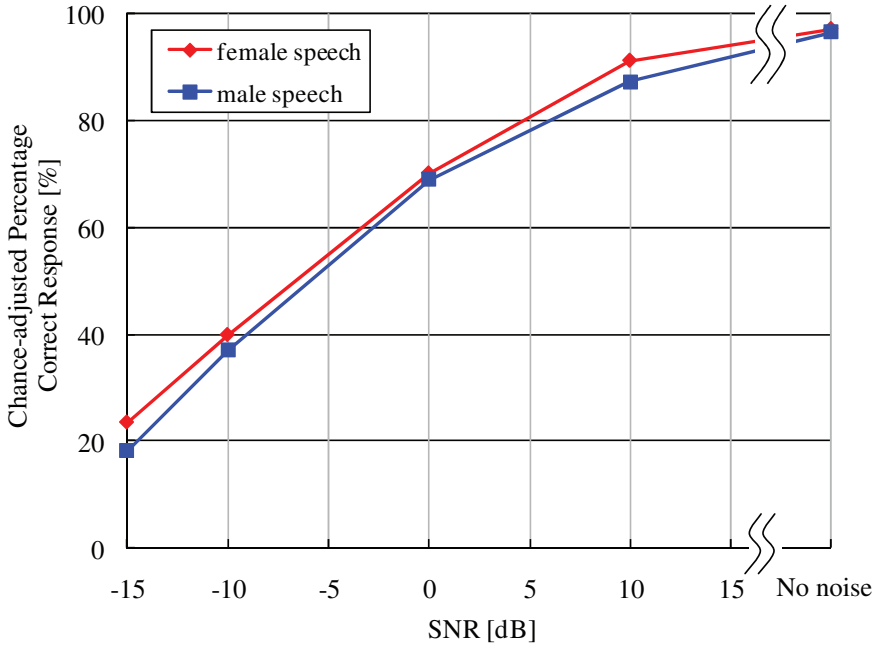

Fig. 5. Comparison of DRT scores of speech mixed with white noise by speaker gender

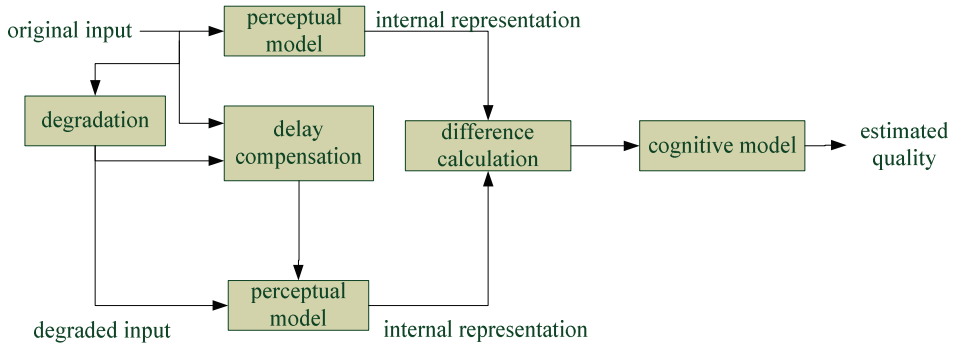

Fig. 6. Simplified diagram of the PESQ algorithm

Figure 7 is the result of an experiment we conducted to estimated the MOS-LQO (Listening Quality Objective), which is an estimated MOS output of the PESQ algorithm (ITU-T, 2003). We used read Japanese sentences of two male and two female speaker, five per speaker for a total of 20 sentences. White noise was added to these speech samples at 30, 10, and -5dB. We also encoded and decoded speech samples with the G.729 CS-ACELP codec (ITU-T, 2007). This codec is commonly used in VOIP applications nowadays. All samples were sampled at 8 kHz, 16 bits per sample. The MOS-LQO 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 this figure, the estimated MOS-LQO generally agrees well with subjective MOS. The line included in the figure is the fitted line using least mean square

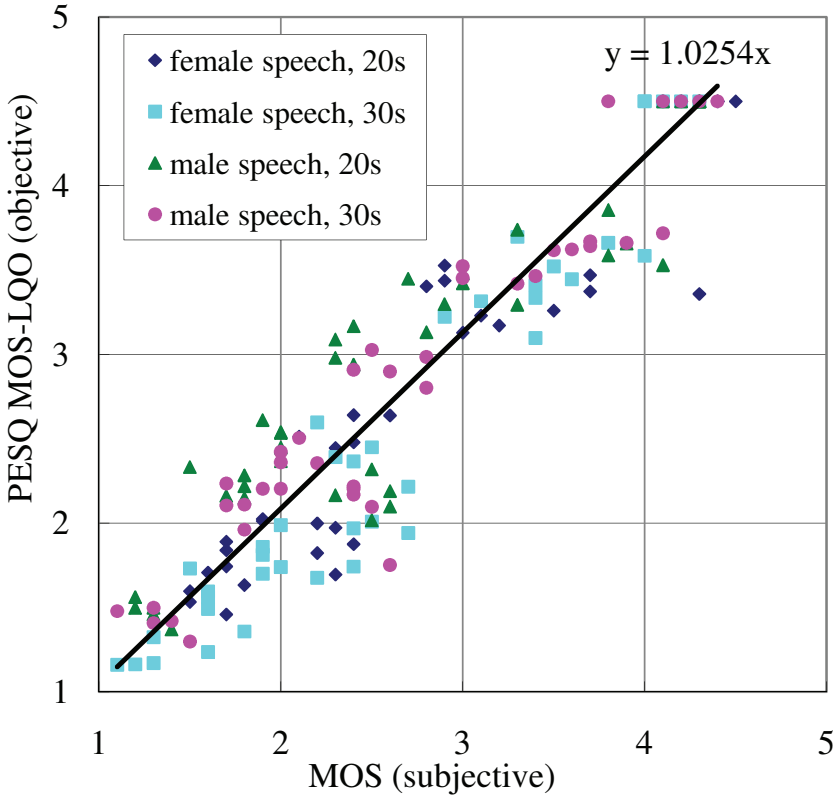

Fig. 7. Example MOS estimation using PESQ

error, which came out to be a gradient of 1.024, also showing that the estimated MOS-LQO generally are accurate estimation of the subjective MOS.

3.3 Correlation between PESQ MOS-LQO and DRT intelligibility score (CACR)

We selected two male and two female speech with standard Japanese DRT words from the collected data described in section 2. All samples were sampled at 16 kHz, 16 bits monaurally. These speech samples were mixed with white noise and babble (Rice University, 1995) at SNR of -15, -10, 0 and 10 dB. Standard subjective DRT tests were run with these samples. Ten listeners were employed. We also estimated MOS-LQO of the degraded samples using PESQ. The wide-band option (+wb) was used in all tests.

Figures 8, and 9 plots the estimated MOS-LQO using PESQ against the corresponding DRT Chance-Adjusted Correct Response (CACR) for speech mixed with white noise, and babble noise, respectively. The speech was pooled for both speakers, for all of the SNR tested in each Figure. As can be seen in both noise types, the correlation between raw MOS-LQO and CACR is quite low. In fact, the Pearson correlation coefficient is 0.47 and 0.44 for female and male speech mixed with white noise, and 0.36 and 0.42 for babble noise. Most of the MOS-LQO is close to the lower end of the scale, *i.e.* well below 2.0, close to 1.0. This is not surprising since PESQ was designed to estimate MOS, and not intelligibility. MOS generally measures

the overall speech quality with relatively small degradation, *i.e.* high SNR range, typically well above 0 dB. However, as we have seen in the previous section, intelligibility is measured in the lower SNR range, typically -20 to 0 dB. Thus we need to re-map the MOS-LQO to match the SNR range of interest for intelligibility estimation.

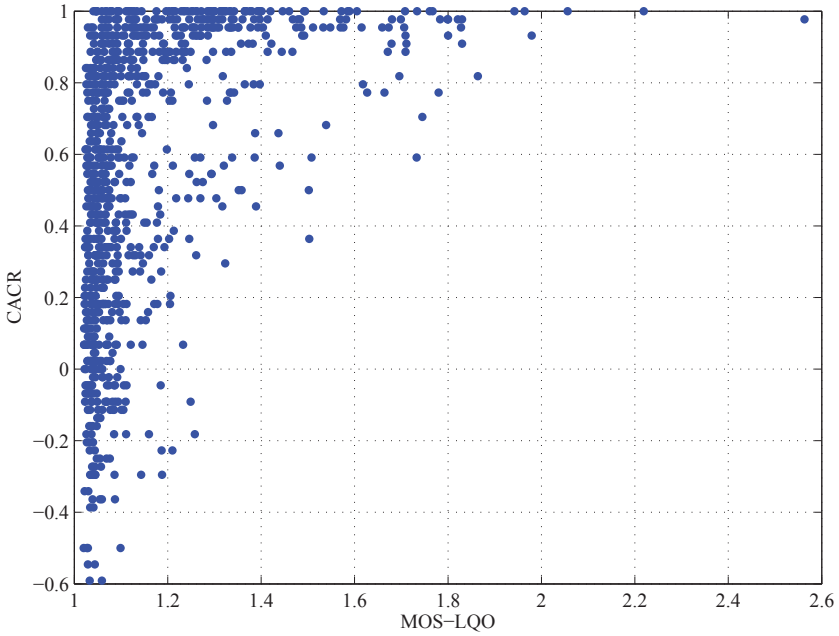

Fig. 8. MOS estimation of DRT words using PESQ (white noise)

3.4 Estimation of intelligibility by mapping per-word MOS-LQO to DRT CACR

We now attempt to map MOS-LQO to CACR using polynomial mapping. We estimated a quadratic polynomial to map the estimated MOS-LQO to DRT CACR on one training speaker. Then we used this polynomial to map MOS-LQO of a different test speaker to DRT CACR. The mapping was estimated for each noise type since it is reasonable to assume that we can obtain a small sample of the noise environment in which we want to estimate the DRT CACR beforehand. We also estimate one polynomial for each phonetic feature, as well as over all features. Table 3 shows an example of the estimated coefficients of the polynomials used to map the female speech mixed with white noise, $y = a_1x^2 + a_2x + a_3$, where x is the MOS-LQO, and y the estimated CACR. As can be seen, the coefficients differ significantly by phonetic feature. The coefficients were also shown to differ significantly by noise type or speaker gender as well.

Tables 4 through 7 tabulate the root mean square DRT CACR estimation error, and the Pearson correlation between subjective and estimated DRT CACR for speech (female and male) mixed with white noise and babble noise, respectively. As can be seen, average estimation errors range from approximately 0.2 to close to 0.7 in some cases. The correlation also ranges from 0.7 to virtually 0.0 in one extreme cases. Thus, the estimation accuracy varies widely by the phonetic feature. Estimation over all features generally perform worse than when using a single phonetic feature.

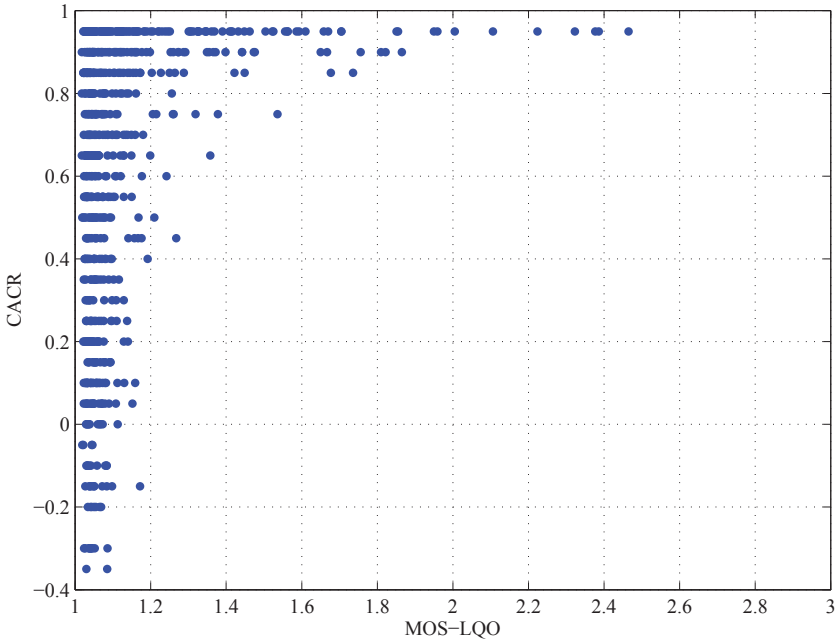

Fig. 9. MOS estimation of DRT words using PESQ (babble noise)

Phonetic feature	a_1	a_2	a_3
voicing	-0.853	2.80	-1.35
nasality	-1.93	6.15	-3.75
sustention	-3.28	10.3	-7.06
sibilation	-1.50	5.15	-3.26
graveness	-1.75	6.02	-4.22
compactness	0.76	-0.893	0.513
all features	-1.84	6.06	-3.94

Table 3. Polynomial coefficients of the mapping function used to map PESQ MOS-LQO to DRT CACR (white noise, female speech)

Figure 10 plots the subjective DRT CACR vs. the estimated DRT CACR for female speech samples for sustention mixed with white noise. This is one of the combinations showing the lowest RMSE and the highest correlation, *i.e.* one of the best predictions. However, the plots scatter widely from the equal rate line. Still the plots are evenly spaced around the equal rate line, and the best fit line is almost equal to the equal rate line. This gives us a clue leading to the approach taken in the next section.

3.5 Estimation of intelligibility by mapping per-feature MOS-LQO to DRT CACR

The standard procedure to measure the subjective intelligibility of a phonetic feature, as measured by CACR, is to test all 20 words on a large listener population, and average

Phonetic feature	RMSE	Correlation
voicing	0.20	0.51
nasality	0.23	0.59
sustention	0.26	0.77
sibilation	0.34	0.54
graveness	0.30	0.63
compactness	0.34	0.49
all features	0.65	0.23

Table 4. Root mean square estimation error and correlation of DRT CACR estimated from PESQ MOS-LQO (white noise, female speech)

Phonetic feature	RMSE	Correlation
voicing	0.68	-0.06
nasality	0.21	0.65
sustention	0.30	0.67
sibilation	0.35	0.52
graveness	0.32	0.61
compactness	0.39	0.41
all features	0.33	0.55

Table 5. Root mean square estimation error and correlation of DRT CACR estimated from PESQ MOS-LQO (white noise, male speech)

Phonetic feature	RMSE	Correlation
voicing	0.27	0.42
nasality	0.33	0.50
sustention	0.32	0.50
sibilation	0.08	0.38
graveness	0.29	0.66
compactness	0.28	0.54
all features	0.52	0.26

Table 6. Root mean square estimation error and correlation of DRT CACR estimated from PESQ MOS-LQO (babble noise, female speech)

the correct response rates for each of the conditions, *e.g.* noise type, SNR, etc. This is because subjective test inherently include a large degree of variations, both due to the tester individuality, and due to variations in the acoustics of the test word speech. By averaging the results for sufficiently large population of testers and over all words in the test list, we can expect to obtain stable reproducible results.

We will attempt the same procedure used to calculate the subjective CACR with the estimated per-word CACR to obtain the per phonetic feature DRT CACR. We pooled all CACR for a

Phonetic feature	RMSE	Correlation
voicing	0.49	0.16
nasality	0.31	0.58
sustention	0.30	0.54
sibilation	0.08	0.36
graveness	0.31	0.61
compactness	0.30	0.56
all features	0.31	0.30

Table 7. Root mean square estimation error and correlation of DRT CACR estimated from PESQ MOS-LQO (babble noise, male speech)

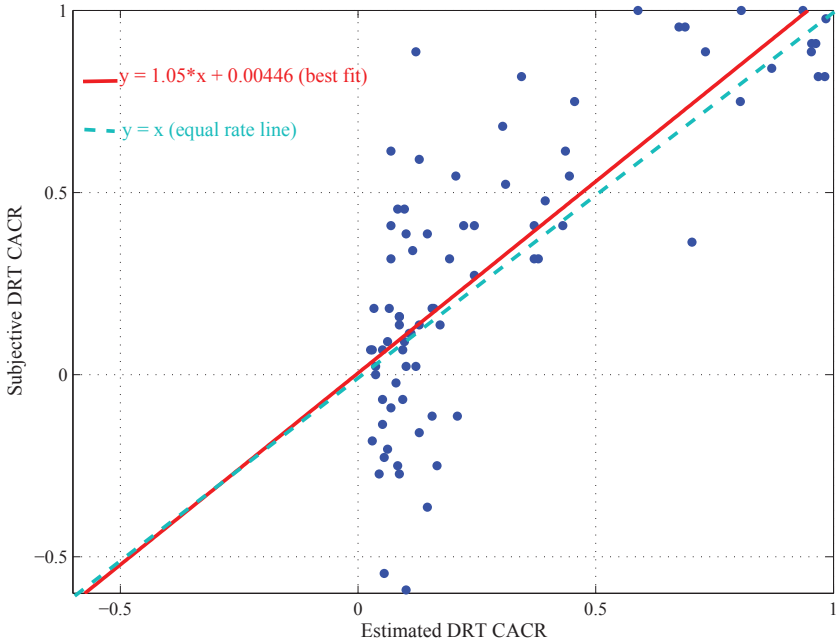

Fig. 10. Subjective CACR vs. estimated CACR (sustention, female speech with white noise)

single phonetic feature, per noise level (SNR) and type, into one CACR. The same quadratic polynomial mapping is used to map the MOS-LQO to DRT CACR, one mapping function per phonetic feature. Again, the mapping was calculated on one training speaker, and this mapping function was used to map MOS-LQO obtained using the PESQ algorithm to calculate the estimated DRT CACR for a different test speaker.

Figures 11 and 12 plot the subjective DRT CACR vs. estimated DRT CACR by pooling for female and male speech in white noise, respectively, while Figures 13 and 14 plot the DRT CACR for female and male speech in babble noise, respectively. Compared to Fig. 10, all plots in these figures are generally much closer to the equal rate line, as expected. This is the result of averaging out the deviation that was present with each of the words in the test word per

phonetic feature. However, as can be seen in Fig. 14, we do not see any estimated DRT CACR below 0.4 for male speech in babble noise. This is due to the limited range that is seen with MOS-LQO under these conditions.

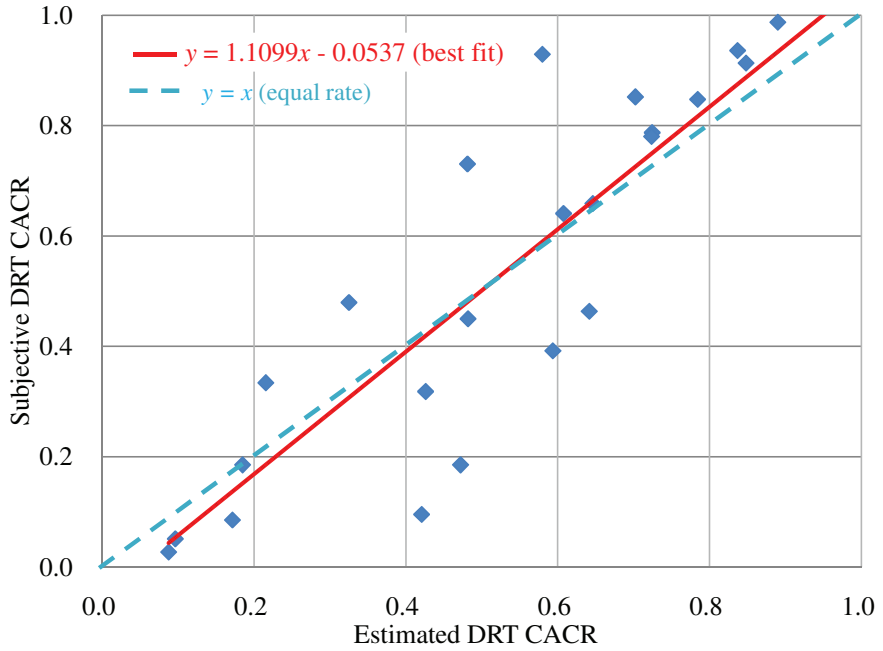

Fig. 11. Subjective CACR vs. estimated CACR (pooled for each feature, female speech with white noise)

Table 8 tabulates the root mean square estimation error and the correlation between subjective and estimated DRT CACR. The RMSE decreased to below 0.2, but even more surprising is the correlation, which is generally above 0.8 now. This level of accuracy is well within practical range if we want to "screen" tested conditions before testing with actual human listeners, as was stated as the goal of this research.

Noise	speaker gender	RMSE	Correlation
white	female	0.15	0.88
white	male	0.20	0.80
babble	female	0.18	0.78
babble	male	0.17	0.82

Table 8. Root mean square estimation error and correlation of DRT CACR estimated from pooled PESQ MOS-LQO

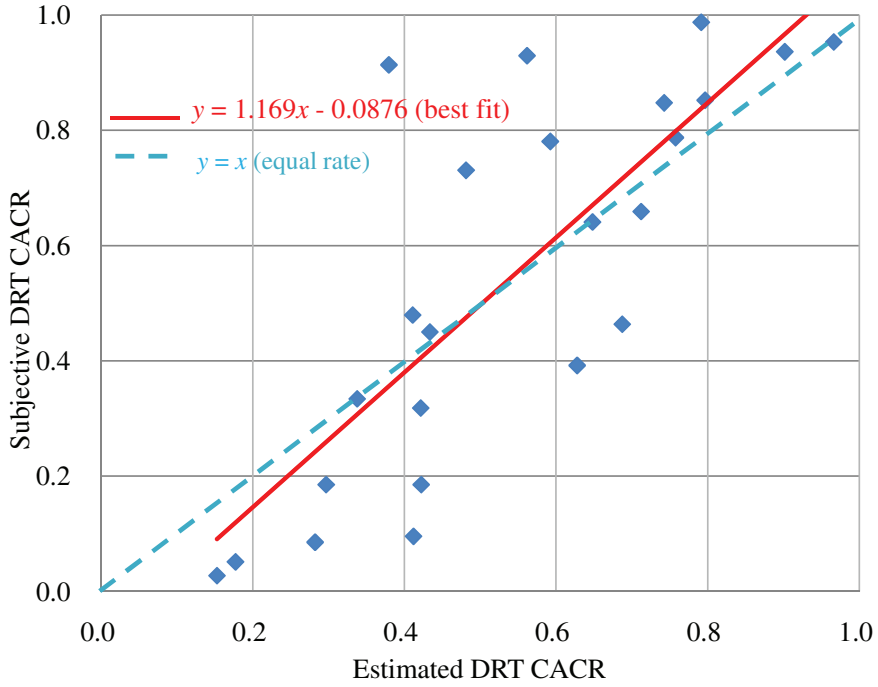

Fig. 12. Subjective CACR vs. estimated CACR (pooled for each feature, male speech with white noise)

4. Conclusion

In this chapter, we have shown that it is possible to estimate the subjective speech intelligibility, as measured by the Diagnostic Rhyme Test (DRT) Chance-Adjusted percentage Correct Rate (CACR), from objective PESQ MOS-LQO scores if we have a mapping function for the noise and the phonetic feature to be tested beforehand. PESQ itself was proven to be too sensitive to noise to serve as a good scale to map to DRT CACR for wide range of signal to noise ratio. In other words, PESQ MOS-LQO saturated quickly to low scores as noise is increased, while DRT CACR stayed relatively high even with considerable noise. This suggests that PESQ may not be a good match to serve as estimation variable for DRT scores for the whole range of SNR we are interested in.

We then attempted to map the MOS-LQO of a test word to DRT CACR using polynomials trained on a training speaker, and mapped the MOS-LQO of an unknown speaker to the DRT CACR. If we use one mapping function per phonetic feature, we showed that it is possible to map the MOS-LQO to DRT CACR to some extent. However, this mapping per word generally showed a large root mean square estimation error (RMSE), mostly larger than 0.3, and the correlation between estimated and subjective DRT CACR was generally low, below 0.5 in most cases.

We then pooled the CACR of the words in each phonetic feature category to estimate the CACR for each feature, as is done in subjective testing. This was shown to dramatically decrease the error, resulting in RMSE below 0.2, and increase the correlation, to above 0.8 in most cases. This dramatic improvement was seen because the estimated CACR for the

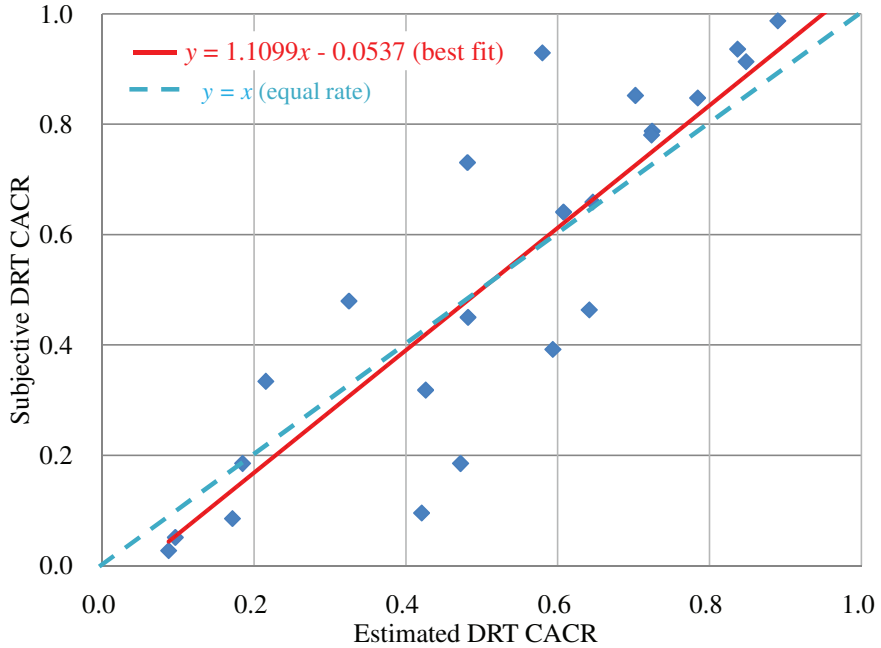

Fig. 13. Subjective CACR vs. estimated CACR (pooled for each feature, female speech with babble noise)

individual words in a phonetic feature category was evenly distributed around the subjective CACR values. By pooling all CACR, we were able to average out this deviation.

Although we have shown that it is possible to estimate the DRT CACR using PESQ-derived MOS-LQO, we still can only do so with limited accuracy. This is because PESQ itself is too sensitive to any amount of noise. Thus, we need to use the internal representation within PESQ and calculate a measure which has more linear correlation with noise levels, or we need to look at completely different objective measures. Accordingly, we have started looking at other candidate objective measures which may show higher correlation with intelligibility, *i.e.* DRT CACR. Segmental SNR and its derivatives, *e.g.* frequency-weighed segmental SNR (Hu & Loizou, 2008), seems to show much higher correlation. The composite measure proposed by the same author, which combines several objective measures, seem to be promising as well (Hu & Loizou, 2008). These measures can be mapped to DRT CACR using polynomials per phonetic feature as we have done in this paper. Preliminary results show significantly improved estimation accuracy. We plan to reveal these results in the near future in a separate paper and conference presentations.

On the other hand, we are also trying out a completely different approach to the same problem. We applied automatic speech recognizers with language models that force one of the words in the word-pair, mimicking the human recognition process of the DRT. The acoustic models were adapted to each of the speakers in the corpus, and then adapted to noise at a specified SNR. We tested with white noise, babble noise, and pseudo-speech noise. The match between subjective and estimated scores improved significantly with noise-adapted models compared to speaker-independent models and the speaker-adapted models, when

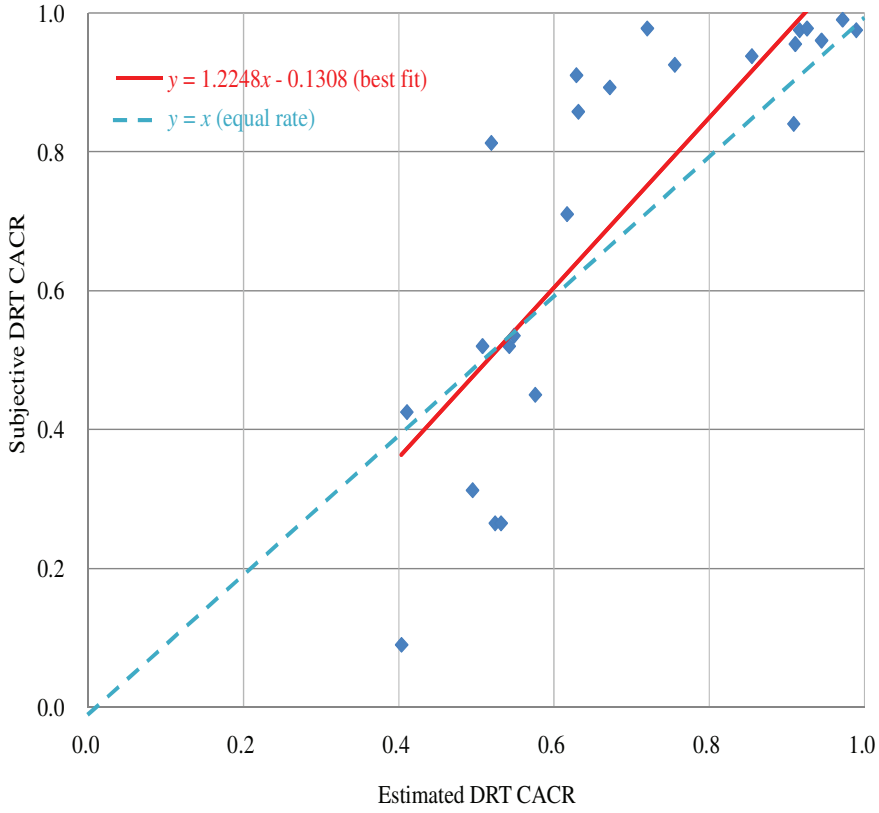

Fig. 14. Subjective CACR vs. estimated CACR (pooled for each feature, male speech with babble noise)

the adapted noise level and the tested level match. However, when SNR conditions do not match, the recognition scores degraded especially when tested SNR conditions were higher than the adapted noise level. Accordingly, we adapted the models to mixed levels of noise, *i.e.*, multi-condition training. The adapted models now showed relatively high intelligibility matching subjective intelligibility performance over all levels of noise. The correlation between subjective and estimated intelligibility scores increased to 0.94 with babble noise, 0.93 with white noise, and 0.89 with pseudo-speech noise, while the root mean square error (RMSE) reduced from more than 0.40 to 0.13, 0.13 and 0.16, respectively. Detailed results are described in a separate paper (Kondo & Takano, 2010; Takano & Kondo, 2010).

5. Acknowledgment

The work described in this chapter was supported in part by the Ministry of Education, Culture, Sports, Science and Technology Grant-in-Aid for Scientific Research (20500151), the Telecommunications Advancement Foundation, and the Research Foundation for the Electro-technology of Chubu. We also thank Professors Tetsuo Kosaka and Masaharu Kato for their assistance in the development of speaker- and noise-adapted speech models. We also

thank Professor Nakagawa as well as our colleagues in the Nakagawa Laboratory for their comments and suggestions. Finally, we thank the students in the Nakagawa Laboratory and Kondo Laboratory for the daily discussions as well as their voluntary participation in the long and strenuous intelligibility tests.

6. References

Amano, S. & Kondo, K. (1999). *Lexical properties of Japanese*, Sanseido, Tokyo. in Japanese. CD publication.

ANSI (1989). Recommendation S3.2-1989: Method for measuring the intelligibility of speech over communication systems.

Beerends, J. G., Buuren, R. V., Vugt, J. V. & Verhave, J. (2009). Objective speech intelligibility measurement basis of natural speech in combination with perceptual modeling, *J. Audio Eng. Soc.* 57(5): 299–308.

Beerends, J. G., Hekstra, A. P., Rix, A. W. & Hollier, M. P. (2002). Perceptual evaluation of speech quality (PESQ) the new ITU standard for end-to-end speech quality assessment part II - psychoacoustic model, *J. Audio Eng. Soc.* 50(10): 765–778.

Fairbanks, G. (1958). Test of phonemic differentiation: The rhyme test, *J. of the Acoustical Society of America* 30: 596–600.

Fujimori, M., Kondo, K., Takano, K. & Nakagawa, K. (2006). On a revised word-pair list for the japanese intelligibility test, *Proc. International Symposium on Frontiers in Speech and Hearing Research*, Tokyo, Japan.

Hu, Y. & Loizou, P. C. (2008). Evaluation of objective quality measures for speech enhancement, *Trans. on Audio, Sp., and Lang. Process.* 16(1): 229–238.

Iida, S. (1987). On the articulation test, *J. of Acoustical Society of Japan* 43(7): 532–536. in Japanese.

ITU-T (1996). ITU-T Recommendation P.800: Method for subjective determination of transmission quality.

ITU-T (2001). 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 .

ITU-T (2003). ITU-T Recommendation P.862.1: Mapping function for transforming P.862 raw result scores to MOS-LQO.

ITU-T (2005). ITU-T Recommendation P.862.2: Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs.

ITU-T (2007). ITU-T Recommendation G.729: Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP).

Jakobson, R., Fant, C. G. M. & Halle, M. (1952). Preliminaries to speech analysis: The distinctive features and their correlates, *Technical Report 13*, Acoustics Laboratory, MIT.

Kaga, R., Kondo, K., Nakagawa, K. & Fujimori, M. (2006). Towards estimation of Japanese intelligibility scores using objective voice quality assessment measures, *Proc. 4th Joint Meeting of the ASA and the ASJ, J. Acoust. Soc. of Am.*, Vol. 120, Honolulu, Hawaii, p. 3255.

Kitawaki, N. & Yamada, T. (2007). Subjective and objective quality assessment for noise reduced speech, *Proc. ETSI Workshop on Speech and Noise in Wideband Communication*, Vol. IV, pp. 1–4.

Kondo, K., Izumi, R., Fujimori, M., Kaga, R. & Nakagawa, K. (2007). On a two-to-one selection based Japanese intelligibility test, *J. Acoust. Soc. of Japan* 63(4): 196–205. in Japanese.

Kondo, K., Izumi, R. & Nakagawa, K. (2001). Towards a robust speech intelligibility test in Japanese, *Proc. 17th International Congress on Acoustics*, Rome, Italy, p. 7P.39.

Kondo, K. & Takano, Y. (2010). Estimation of two-to-one forced selection intelligibility scores by speech recognizers using noise-adapted models, *Proc. Interspeech*, ISCA, Makuhari, Japan, pp. 302–305.

Liu, W. M., Jellyman, K. A., Evans, N. W. D. & Mason, J. S. D. (2008). Assessment of objective quality measures for speech intelligibility, *Proc. Interspeech*, Brisbane, Australia.

Miller, G. A., Heise, G. A. & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials, *J. of Experimental Psychology* 41: 329–335.

NHK Broadcasting Culture Research Institute (ed.) (1998). *Japanese Pronunciation Dictionary*, Japan Broadcast Publishing.

Nishimura, R., Asano, F., Suzuki, Y. & Sone, T. (1996). Speech enhancement using spectral subtraction with wavelet transform, *IEICE Trans. Fundamentals* 79-A(12): 1986–1993. in Japanese.

Rice University (1995). Signal Processing Information Base (SPIB), `http://spib.rice.edu/spib/select_noise.html`.

Rix, A. W., Hollier, M. P., Hekstra, A. P. & Beerends, J. G. (2002). Perceptual evaluation of speech quality (PESQ) the new ITU standard for end-to-end speech quality assessment part I - time-delay compensation, *J. Audio Eng. Soc.* 50(10): 755–764.

Suzuki, Y., Kondo, K., Sakamoto, S., Amano, S., Ozawa, K. & Sone, T. (1998). Perceptual tendency in word intelligibility tests by use of word-lists with controlled word familiarities, *Technical Report H-98-47*, Acoustical Society of Japan Technical Committee on Psychological and Physiological Acoustics. in Japanese.

Takano, Y. & Kondo, K. (2010). Estimation of speech intelligibility using speech recognition systems, *IEICE Trans. on Inf. and Syst.* E93-D(12): 3368–3376.

Tanaka, M. (1989). A prototype of a quality evaluation system for hearing aids, *Technical report*, Report of the Results of Research with METI Kakenhi (Grant-in-Aid). in Japanese.

Voiers, W. D. (1977). *Speech Intelligibility and Speaker Recognition*, Dowden, Hutchinson & Ross, Stroudsburg, PA, chapter Diagnostic Evaluation of Speech Intelligibility, pp. 374–387.

Voiers, W. D. (1983). Evaluating processed speech using the diagnostic rhyme test, *Speech Technology* 1: 30–39.

Speech and Language Technologies

Edited by Prof. Ivo Ipsic

ISBN 978-953-307-322-4
Hard cover, 344 pages
Publisher InTech
Published online 21, June, 2011
Published in print edition June, 2011

This book addresses state-of-the-art systems and achievements in various topics in the research field of speech and language technologies. Book chapters are organized in different sections covering diverse problems, which have to be solved in speech recognition and language understanding systems. In the first section machine translation systems based on large parallel corpora using rule-based and statistical-based translation methods are presented. The third chapter presents work on real time two way speech-to-speech translation systems. In the second section two papers explore the use of speech technologies in language learning. The third section presents a work on language modeling used for speech recognition. The chapters in section Text-to-speech systems and emotional speech describe corpus-based speech synthesis and highlight the importance of speech prosody in speech recognition. In the fifth section the problem of speaker diarization is addressed. The last section presents various topics in speech technology applications like audio-visual speech recognition and lip reading systems.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821