Applied Acoustics 106 (2016) 113–121

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Estimation of forced-selection word intelligibility by comparing objective distances between candidates

Kazuhiro Kondo

Graduate School of Science and Engineering, Yamagata University, 4-3-16 Jonan, Yonezawa, Yamagata 9928510, Japan

ARTICLE INFO

Article history:
Received 15 September 2015
Received in revised form 11 December 2015
Accepted 4 January 2016
Available online 18 January 2016

Keywords:
Speech intelligibility
Objective measures
Forced selection test
Mapping

ABSTRACT

We proposed and evaluated an estimation method for the forced selection speech intelligibility tests. Our proposal takes into account the forced selection manner of the Diagnostic Rhyme Test (DRT), which forces selection from a pair of rhyming words. A distance measure is calculated between the test word and the two candidate words, respectively, and the distance is compared to select the most likely word. We compared two distance measures. The first objective distance measure used here was based on the Articulation index Band Correlation (ABC). The ABC is the correlation of time–frequency (T–F) patterns between the test word and the template word speech of the two words in the candidate word pair. The word with the higher correlation was decided to be the likely candidate word. The T–F pattern was calculated in the Articulation Index (AI) bands, and the correlation was calculated between the corresponding bands of the test and candidate word sample. In order to estimate the intelligibility, we calculate the ratio of the number of bands in which higher correlation is seen for the correct word vs. the total number of bands (named ABC-est). This ratio quantifies how well the test word matches the correct word in the word pair. For the second objective distance, we used a measure based on the frequency-weighted segmental SNR (fwSNR$_{seg}$). Segmental SNR (SNR$_{seg}$) was calculated in AI bands, and compared among the candidate word templates. We then calculated the frequency-weighted ratio of the number of bands in which higher SNR$_{seg}$ was observed for the correct word vs. the total number of bands (named fwSNR$_{seg}$-est), again to quantify how well the test word matches the selected candidate word in the pair. We estimated a logistic mapping function from the above two ratios to intelligibility scores using speech mixed with known noise. The mapping functions were then used to estimate the intelligibility of speech mixed with unknown noise. This estimation was compared to another measure that we previously evaluated, the conventional fwSNR$_{seg}$, which directly maps the measure to intelligibility. Both proposed measures were proven to be significantly more accurate than conventional fwSNR$_{seg}$. For most cases, the accuracy was comparable between the two proposed distance measures, ABC-est and fwSNR$_{seg}$-est, with the latter showing correlation between the subjective and estimated intelligibility as high as 0.97, and root mean square as low as 0.11 for one of the test sets, but not as accurate for other sets. The ABC-est showed more stable accuracy for all sets. However, both measures show practical accuracies in all conditions tested. Thus, it should be possible to "screen" the intelligibility in many of the noise conditions to be tested, and cut down on the scale of the subjective test needed.

1. Introduction

With the advanced mobile communication networks becoming ubiquitous, speech communication is taking place in a variety of ambient noise environments. Thus, comprehensive speech communication quality measurement techniques are required, and frequent evaluation of this quality is needed.

There are basically two forms of speech quality measures; the overall perceptual listening quality, and the speech intelligibility [1,2], which measures the accuracy of the perceived speech signal. Of these, we will deal with the latter here.

Japanese intelligibility tests have often used randomly selected single mora, two or three morae words as stimuli. The subjects were asked to choose freely from any combination of valid Japanese syllables. This selection task quickly becomes a difficult one as the added degradation to the stimuli increase.

The situation is similar for English tests. Thus, for English intelligibility tests, closed set selection tests that restricted the

http://dx.doi.org/10.1016/j.apacoust.2016.01.003
0003-682X/

response to a choice from a limited number of candidates were introduced. For example, the Modified Rhyme Test (MRT) allows the subjects to choose from a list of six words [3], while the Diagnostic Rhyme Test (DRT) uses two word candidates [4,5]. These types of tests are said to be effective in controlling various factors, including the amount of training and phonetic context. We have previously shown that the Japanese DRT gives stable intelligibility scores, even with a relatively naive panel of listeners [6]. Because of the simplicity of the structure of this test, its administration can easily be automated using computers, and the data processing to calculate the accuracy is also simple and can be fully automated as well. The English DRT has now become an ANSI standard [7], and is widely accepted as one of the intelligibility measurement tests.

We have developed and proposed a DRT test for Japanese [6,8]. This test has the same structure as its English counterpart, and was shown to give stable results for additive noise degradations.

Even with these forced selection tests, evaluations using human subjects are expensive and time-consuming. Accordingly, numerous efforts to estimate intelligibility without humans have been conducted. For example, AI, proposed by French and Steinberg [9], estimates the intelligibility from SNR measurements within several frequency bands combined using a perceptual model. This evolves to a number of measures, including the Speech Transmission Index (STI) [10] by Steeneken and Houtgast, which uses artificial speech communicated over the channel to estimate the intelligibility by measuring the modulation depth of weighted frequency bands of the received signal. Recently, Schwerin and Paliwal improved the accuracy of STI for degraded speech with non-linear distortion by introducing short analysis segments to compensate for the quasi-stationary nature of speech [11], which they named the Quasi-stationary STI (QSTI). Ma et al. have shown that by using a new signal-dependent time-varying band importance functions (BIFs) on conventional objective measures, such as the Signal to Noise Ratio (SNR), AI-based measures, and others, the estimation accuracy can be improved [12]. Taal et al. introduced a simple objective measure, which they call the Short-Time Objective Intelligibility (STOI) measure [13]. The STOI measures the correlation between the temporal envelopes of clean and degraded speech in short segments. They have shown that speech intelligibility can be estimated much more accurately than previous methods. Recently, Unoki et al. have been attempting to estimate the STI of speech with reverberation using a model that simulates the effect of reverberation on the modulation depth, and showed promising results [14].

We also previously showed that it is possible to estimate the intelligibility measured using the DRT by mapping objective scores to the intelligibility using a pre-trained regression function. Some objective scores we have tried were the Mean Opinion Scores (MOS) calculated using the Perceptual Evaluation of Speech Quality (PESQ) [15,16], and $fwSNR_{seg}$ [17,18]. Estimation using these objective measures proved to be relatively accurate provided that a sampling of the noise is available to train the regression models.

In this paper, we attempt to further implement an estimation method that takes into account the forced selection manner of the DRT from a pair of rhyming words. The objective measure used here was based on ABC proposed by Voran [19]. Voran showed that relatively high estimation accuracy can be achieved with ABC for the English MRT [3].

We applied this ABC to DRT and calculated the correlation of T–F patterns between the test word and the template word speech of the two words in the candidate word pair. The word with the higher correlation is chosen as the most likely candidate word. The T–F pattern is calculated by frequency bands used in the AI standard, and correlation is calculated between the corresponding bands. The candidate word with more AI bands showing higher correlation values is chosen.

We use the above ABC to calculate the distance between the correct word template and the test word as well as the incorrect word in the pair, respectively. In order to do so, the ratio of the number of bands in which higher correlation can be observed between the correct word template and the test speech vs. the total number of bands is calculated in order to quantify how well the test word matches the correct word in the word pair.

In the above estimation, the calculated distance measure used to discriminate between the candidate words was the T–F pattern. However, we previously showed that $fwSNR_{seg}$ itself is an extremely robust distance measure [18]. Accordingly, we will also employ a distance based on this measure in order to discriminate the match between the candidate words in the word-pair. As was done with the T–F pattern, we will measure the distance between the test sample and candidate words using the segmental SNR (SNR_{seg}) calculated in each AI band, and use the frequency-weighted ratio of the number of bands showing smaller distance between the correct candidate word vs. the total number of bands, again to quantify the degree of the match between the test and candidate word.

With both of the above distance measures, the ratios are used to train regression functions between subjectively measured intelligibility, and then used to estimate the intelligibility of an unknown condition from the calculated distance measure. As will be described in this paper, both of these objective measures show high estimation accuracy.

This paper is organized as follows. In the next section, the estimation methods for forced-selection word intelligibility using ABC, and the $fwSNR_{seg}$ is described. In Section 3, the estimation accuracy evaluations of the proposed methods are described. This is followed by the results and its analysis in Section 4. Finally, the conclusions and suggestions for further work are given.

2. Estimation of forced-selection word intelligibility using objective measures

We explicitly compared the distance between the noisy speech under test to both of the candidate word speech in the word pair, and measured the difference in the distance between the "correct" candidate word (which we assume is known during the estimation process) and the competing word in the pair according to the various noise types and SNRs. This distance difference was averaged and used to estimate the subjective intelligibility scores of this forced-selection intelligibility test using regression.

We compared two types of distances. The first was based on the ABC proposed by Voran [19]. The other was the segmental SNR calculated in AI bands. These distance measures will be described in detail in the following sections.

2.1. The Japanese diagnostic rhyme test

In this paper, the subjective speech intelligibility was measured using the Japanese DRT [6,8]. DRT is a speech intelligibility test 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. The features used in the DRT, following the definition by Jacobson et al. [20], are voicing, nasality, sustention, sibilation, graveness, and compactness.

A brief description of the definition of the features along with an example word-pair in the Japanese DRT is shown in Table 1.

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 [8]. The word-pairs are rhyming words, differing only in the initial consonant. One of the words in the word-pair list is a word whose initial consonant has the consonant feature under test, and the initial consonants in the other word does not.