ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Speech Communication
Session 1aSCb: Digital Speech Processing (Poster Session)

1aSCb13. Relationship between subjective and objective evaluation of noise-reduced speech with various widths of temporal windows
Mitsunori Mizumachi*

*Corresponding author's address: Dept. of Electrical Engineering and Electronics, Kyushu Institute of Technology, 1-1 Sensui-cho, Tobata-ku, Kitakyushu, 804-8550, Fukuoka, Japan, mizumach@ecs.kyutech.ac.jp

It is necessary to enhance adverse speech signals for building useful speech interfaces. Speech enhancement is essential under noisy and reverberant acoustic environments. Therefore, quality assessment of the enhanced speech signals should be also an important issue in noise reduction and dereverberation. Subjective evaluation is given by carrying out listening tests, and objective evaluation is provided by speech distortion measures. However, there is the discrepancy between subjective and objective evaluation of speech distortion. The author has investigated the relationship between subjective and objective evaluation of noise-reduced speech signals. The objective speech distortion was calculated in each short-term frame, of which length was fixed, and the statistical characteristics of the short-term speech distortion were investigated using higher-order statistics such as skewness and kurtosis. The preliminary result suggested that skewness of the short-term speech distortion could give an explanation for the discrepancy between subjective and objective evaluation. Further investigation of the relationship between subjective and objective evaluation of noise-reduced speech signals is carried out with a variety of temporal window widths. [Work supported by NEDO, Japan]
INTRODUCTION

It is important to provide high quality speech signals for speech applications. In the real world, however, observed speech signals are often distorted by background noises and room reverberation. It is indispensable for the speech interface to equip the function of noise reduction. It is also important to evaluate the performance of noise reduction properly. It should be considered both how much the interfering noise is reduced and how the target signal is preserved [1]. A noise-reduced speech signal is assessed by means of employing an objective distortion measure or carrying out a subjective listening test. Practically, it is impractical to carry out the listening test on the stage of designing the noise reduction algorithm frequently. It is advisable to prepare a reasonable objective distortion measure.

Objective distortion measures are divided into signal-based methods and knowledge-based methods. The signal-based distortion measures include the signal-to-noise ratio (SNR), of which feature is energy in the temporal domain, the spectral distance (SD), which calculates the Euclidean distance between amplitude spectra, and the cepstral distance (CD), which calculates the Euclidean distance between spectral envelopes. On the other hand, there are a wide variety of the knowledge-based distortion measures, which imitate our auditory characteristics. The well-known distortion measures include the perceptual evaluation of speech quality (PEAQ) [2] and the perceptual evaluation of audio quality (PEAQ) [3] for narrow-band and wide-band speech signals, respectively. These distortion measures are often employed to evaluate noise reduction algorithms, although they are originally designed for evaluating slight distortion caused by speech and audio codecs. Then, we should not use the PESQ and the PEAQ for evaluating the noise-reduced speech signals under the low SNR conditions. Indeed, both the PESQ and the PEAQ are incompatible with our auditory impression on heavily distorted speech signals.

This paper focuses on the relationship between the speech distortion in the short-term segment and the ease of speech perception. As the primitive objective distortion measure, the segmental SNR is calculated in each short-term frame, and the variation of the segmental SNR is simply described using the descriptive statistics. On the other hands, the subjective evaluation is given on the 5-grade mean opinion score (MOS) by a listening test. In addition to the mean of the segmental SNR, higher order statistics including variance, skewness, kurtosis, are employed to quantify the temporal variation of the segmental SNR. Those relationships among the MOS and the descriptive statistics are investigated with the different frame length in calculating the segmental SNR.

DATA PREPARATION

In this paper, the relationship between objective and subjective evaluation is investigated under the severe noisy conditions. The relationship under the low SNR conditions has not been discussed enough, while the relationship is well investigated under the high SNR conditions. Then, the heavily distorted speech signals are prepared, of which SNRs are less than 0 dB.

Preparation of Noisy Speech

The isolated 7-digit utterance by the female speaker is prepared for the target speech signal from the TI-digit speech database [4]. When we have any difficulty in speech perception, we can predict the missing speech segment within the single digit. However, the prediction is difficult across multiple digits. Therefore, the isolated digit utterance is suitable for the target speech under the severe noisy condition. The target speech was played through a loud speaker and re-recorded in the soundproof room, of which reverberation time is 0.3 s.

Noise data are prepared from the NOISEX-92 database [5]. Five different characteristics of noise signals are carefully selected as follows. The station yard noise and the exhibition hall noise have the similar characteristics to speech in the frequency domain, but the temporal characteristics of them are different from speech. Three types of the non-stationary factory floor noises are selected out, which have different spectral and temporal characteristics mutually. Those noise signals were also re-recorded in the soundproof room.

The target speech and interfering noise signals were re-recorded independently, and were mixed in a computer at the desired SNR condition, where the SNR is set at the four steps under 0 dB in SNR. In total, 20 kinds of noisy speech are prepared in 5 noise types and 4 SNR conditions.
Preparation of Noise-reduced Speech

Noisy speech signals are enhanced by the linear and non-linear noise reduction schemes. Acoustical characteristics among the noise-reduced speech signals are mutually different by the linear and non-linear signal processing.

The linear noise reduction is achieved by 8-ch delay-and-sum beamforming, where the amount of noise reduction is less but no distortion is occurred on the target speech [6]. To obtain the 8-ch noisy observation, each noise source played by the loudspeaker was re-recorded using the 8-ch equally-spaced linear microphone array. The direction of the arrival target speech signal was exactly given in beamforming. The alternative non-linear method is the combination of the 8-ch delay-and-sum beamforming with the spectral subtraction [7]. The spectral subtraction is widely used for the practical speech interfaces such as mobile phones due to its high efficiency, but causes the annoying distortion on the post-processed speech signal. The amplitude spectrum of each noise signal after beamforming was trained in advance for the spectral subtraction.

RELATIONSHIP BETWEEN OBJECTIVE AND SUBJECTIVE EVALUATION

The noisy and noise-reduced speech samples are evaluated objectively and subjectively. In this section, the relationship between objective and subjective evaluation is confirmed globally and locally.

Objective Evaluation

In this paper, the short-term segmental SNR is employed as the most primitive distortion measure in order to consider the temporal variation of speech distortion. In this Section, SNR is calculated using the short-term speech segment of 21.3 ms, where 1,024 samples are in each segment with 48 kHz/16 bits accuracy, and the frame was shifted by 10.6 ms. The segmental SNR is calculated only in the speech period, of which boundaries are manually determined.

Subjective Evaluation

Listening test was carried out to evaluate the noisy and the noise-reduced speech samples with 25 participants, who were graduates and undergraduates with normal hearing. Subjective impression for the distorted speech signal varies depending on the noise type, the SNR, and the interaction between speech and noise. In this paper, each speech sample was evaluated on the 5-grade mean opinion score (MOS) concerning the ease of speech perception in the range: (5) can understand definitely, (4) can understand probably, (3) can understand somehow, (2) cannot understand partially, and (1) cannot understand at all. The 20 noisy speech samples and the 40 noise-reduced speech samples were randomly presented five times, and then each participant gave the MOS on the ease of speech perception in 300 trials. Concerning the five MOSs given by each subject per each speech sample, the four scores except the score given at the first turn were averaged over. Finally, the participant-depended MOS was averaged over 25 participants.

Relationship Between Subjective and Objective Evaluation

It is well known that the SNR, which is usually calculated for the whole utterance, does not correspond to the MOS. To confirm the problem, the short-term segmental SNR are averaged over within each utterance as usual, and the global relationship between the mean SNR and the MOS is given in Fig. 1. For each of the noisy and the noise-reduced speech conditions, the 5 noise types and the 4 SNR conditions are averaged over. Concerning the global SNR, 2.6 dB and 4.7 dB improvement in SNR were obtained for the beamforming and the beamforming with the spectral subtraction, respectively. However, the MOSs are not proportional to the global SNR. This result also confirms the discrepancy between subjective and objective evaluation.
In this section, the detailed relationship between objective and subjective evaluation is investigated to discover the factor of the above discrepancy. Figure 2 shows the relationship between the mean SNR and the MOS for each of the 60 speech samples. In Fig. 2, the noisy speech, the noise-reduced speech only by the beamforming, and the noise-reduced speech by the beamforming with the spectral subtraction are represented with asterisk, circle, and square marks, respectively. It is confirmed that the MOSs at the same SNR differ between the noise types and the signal processing conditions. In short, the mean of the segmental SNR is not feasible to predict the subjective impression of noisy speech.

**Analysis with Different Length of Temporal Window**

The local relationship between objective and subjective evaluation is further investigated under the different analysis conditions. The short-term SNR can consider the temporal variation of speech distortion. In Fig. 3, the relationships between objective and subjective evaluation are given with the frame length of 5.3 ms, 85.3 ms, and 1370 ms, respectively. The aspect of the correlation between the mean SNR and the MOS changes depending on the frame length. The discrepancy between objective and subjective evaluation is smaller with shorter frame length.
FIGURE 3. Relationship between the mean SNR and the MOS under the different frame lengths of 5.3 ms, 85.3 ms, and 1370 ms, respectively. In each panel, the noisy speech, the noise-reduced speech only by the beamforming, and the noise-reduced speech by the beamforming with the spectral subtraction are represented with asterisk, circle, and square marks, respectively.
STATISTICAL ANALYSIS OF SPEECH DISTORTION

In this section, we account for the discrepancy between objective and subjective evaluation of distorted speech. The author has confirmed that the statistical property of the segmental SNRs in each utterance might explain the reason of the discrepancy between subjective and objective evaluation [8]. It is the key to quantify the unevenness of the segmental SNR in the short range of the MOS.

Formulation of Statistical Variation of Segmental SNR

To simply describe the characteristics of the segmental SNR distribution, the descriptive statistics are employed. The mean is the 1st order descriptive statistic. It is, however, confirmed that the mean of the segmental SNR is not sufficient for explaining the ease of speech perception. In this paper, the higher-order statistics are employed such as the variance, the skewness, and the kurtosis. The skewness is the third order moment of the distribution, and represents the symmetricalness around the mean as follows:

\[
Skewness(x) = \frac{1}{N} \sum_{i=1}^{N} (x - \bar{x})^3 \frac{1}{\left( \frac{1}{N} \sum_{i=1}^{N} (x - \bar{x})^2 \right)^{3/2}},
\]

where \( x \) means the short-term segmental SNR, \( \bar{x} \) is the mean of the segmental SNR, and \( N \) is the number of the segmental SNRs. Note that the denominator is the third power of the standard deviation. Next, the kurtosis is the fourth order moment of the distribution, and represents the sharpness around the mean as follows:

\[
Kurtosis(x) = \frac{1}{N} \sum_{i=1}^{N} (x - \bar{x})^4 \frac{1}{\left( \frac{1}{N} \sum_{i=1}^{N} (x - \bar{x})^2 \right)^{2}}.
\]

Analysis of Segmental SNR Distribution Based on Descriptive Statistics

The descriptive statistics are calculated to quantify the deviation of speech distribution. In Fig. 3, the discrepancy between the mean SNR and the MOS is remarkable in the rage of 2 to 3 in MOS not depending on the frame length. Therefore, only the speech samples evaluated in the above MOS range are used to understand the discrepancy problem. The noisy speech samples are compared with the noise-reduced speech samples by the beamforming method, because those two conditions bring the difference in the relationship between the mean SNR and the MOS. The descriptive statistics for the selected speech samples in each frame length are summarized in Table 1.

<table>
<thead>
<tr>
<th>Frame length [ms]</th>
<th>Condition</th>
<th>Mean MOS</th>
<th>Segmental SNR [dB]</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>mean</td>
<td>S.D.</td>
<td>skewness</td>
<td>kurtosis</td>
</tr>
<tr>
<td>5.3</td>
<td>Noisy</td>
<td>2.15</td>
<td>-5.65</td>
<td>0.43</td>
<td>-0.29</td>
<td>1.80</td>
</tr>
<tr>
<td></td>
<td>Noise-reduced</td>
<td>2.19</td>
<td>-7.28</td>
<td>0.45</td>
<td>-0.39</td>
<td>1.94</td>
</tr>
<tr>
<td>85.3</td>
<td>Noisy</td>
<td>2.15</td>
<td>-7.56</td>
<td>4.35</td>
<td>-0.96</td>
<td>2.85</td>
</tr>
<tr>
<td></td>
<td>Noise-reduced</td>
<td>2.19</td>
<td>-9.24</td>
<td>4.39</td>
<td>-0.86</td>
<td>2.74</td>
</tr>
<tr>
<td>1370</td>
<td>Noisy</td>
<td>2.15</td>
<td>-5.89</td>
<td>2.97</td>
<td>-0.74</td>
<td>2.91</td>
</tr>
<tr>
<td></td>
<td>Noise-reduced</td>
<td>2.19</td>
<td>-7.53</td>
<td>3.12</td>
<td>-0.53</td>
<td>2.38</td>
</tr>
</tbody>
</table>
Discussion

It is confirmed that the averaged MOSs are the almost the same in between the noisy and the noise-reduced conditions. The mean MOSs are the same in all frame lengths, because they are originated with the subjective evaluation. The means of the segmental SNRs are significantly different in between the noisy and the noise-reduced conditions as shown in Fig. 3.

The standard deviations were extremely close in between the two conditions. Therefore, the standard deviation of the segmental SNRs could not explain the discrepancy between the mean SNR and the MOS. The kurtosis was also not distinctive. The skewness gave the different tendency in between the two conditions, and represented the bias of the segmental SNR. The symmetricalness of the segmental SNR distribution could be one of the important features in evaluating speech distortion. It is important to use both the mean and the skewness of the segmental SNRs in order to predict the subjective impression of noisy and noise-reduced distorted speech.

CONCLUSIONS

This paper reviewed the discrepancy between objective and subjective evaluation in speech distortion. To discover the factor of the discrepancy, the statistical characteristics of the short-term segmental SNR are analyzed with higher order descriptive statistics. It was confirmed that both the mean and the skewness of the segmental SNR are important features in predicting the ease of speech perception. It is expected that this findings contribute to design an appropriate speech distortion measure in future.

ACKNOWLEDGMENTS

This research was partially supported by the New Energy and Industrial Technology Development Organization (NEDO) of Japan.

REFERENCES