Speech recognition performance estimation for clipped speech based on objective measures

Yuuki Tachioka*, Tomohiro Narita† and Jun Ishii‡

Information Technology R & D Center, Mitsubishi Electric Corporation,
5-1-1 Ofuna, Kamakura, 247-8501 Japan

(Received 22 April 2014, Accepted for publication 16 June 2014)

Keywords: Clipping, Laplacian, SNR, PESQ, Speech recognition performance, Confusion matrix
PACS number: 43.60.Cg, 43.60.Dh, 43.72.Ne [doi:10.1250/ast.35.324]

1. Introduction

Recently, speech recognition systems have been used in various environments. As a result, we are facing low-quality speech data. One of the factors that degrade speech quality is clipping mainly caused by inappropriate microphone settings. Apparently, clipping degrades speech quality, and a study [1] has shown that speech recognition performance deteriorates for real data. However, in that study [1], other factors, such as contents of utterances, recording devices or speakers, are uncontrolled and the influence of clipping itself on speech recognition performance cannot be evaluated quantitatively. In this study, we clarify the relationship between clipping level, which represents the extent of clipping, and automatic speech recognition (ASR) performance for various artificially clipped utterances by changing the clipping level but keeping the other factors the same.

Practically, ASR performance estimation is useful. There are some studies on estimating the ASR performance in reverberant and noisy environments [2,3] but, to the best of our knowledge, there are as yet no studies on estimating the ASR performance of clipped speech. Our experiments show that the ASR performance can be expressed as a logistic regression using the signal-to-noise ratio (SNR) and perceptual evaluation of speech quality (PESQ) [4]. Moreover, we show the explicit relationship between SNR and clipping level, theoretically.

2. Definition of clipped signals and clipping level

Clipped signals $\hat{y}$ are obtained by clipping original signals y with a clipping level θ_c , as shown in Eq. (1), after y is normalized between -1 and 1 .

$$\hat{y} = \begin{cases} \text{sign}(y)\theta_c & (\text{if } |y| \geq \theta_c) \\ y & (\text{otherwise}) \end{cases} \quad (1)$$

Here, “sign” returns to 1 if the argument is positive or to -1 otherwise. When θ_c is one, $\hat{y}$ equals y .

3. SNR estimation of clipped speech

For clipped speech, SNR ψ is a function of θ_c ,

$$\psi(\theta_c) = 10 \log_{10} \frac{\sum_{t \in \mathcal{T}_c} y_t^2}{\sum_{t \in \mathcal{T}_c} (y_t - \hat{y}_t)^2}, \quad (2)$$

where t is the sample number and $\mathcal{T}_c$ is a set of the indexes of clipped samples.

In accordance with Ref. [5], we assume that a background model of speech is represented as a Laplacian,

$$p(y) = \frac{1}{2b} \exp\left(-\frac{|y|}{b}\right), \quad (3)$$

where $2b^2$ is the variance of the Laplacian. The expectation of the numerator of Eq. (2) is given as

$$\begin{aligned} \sum_{t \in \mathcal{T}_c} y_t^2 / N_c &= \int_{\theta_c}^1 y^2 p(y) dy + \int_{-1}^{-\theta_c} y^2 p(y) dy, \\ &= 2 \int_{\theta_c}^1 y^2 p(y) dy, \\ &= ((\theta_c + b)^2 + b^2) e^{-\frac{\theta_c}{b}} - (b_1^2 + b^2) e^{-\frac{1}{b}}, \end{aligned} \quad (4)$$

where N_c is the number of clipped samples and b_1 is $1 + b$. Here, we use the relations below:

$$\begin{aligned} f_1 &= \int_{\theta_c}^1 p(y) dy = \frac{1}{2} (e^{-\frac{\theta_c}{b}} - e^{-\frac{1}{b}}), \\ f_2 &= \int_{\theta_c}^1 y p(y) dy = \frac{1}{2} (\theta_c e^{-\frac{\theta_c}{b}} - e^{-\frac{1}{b}}) + b f_1, \\ f_3 &= \int_{\theta_c}^1 y^2 p(y) dy = \frac{1}{2} (\theta_c^2 e^{-\frac{\theta_c}{b}} - e^{-\frac{1}{b}}) + 2b f_2. \end{aligned}$$

On the other hand, the expectation of the denominator is given as

$$\begin{aligned} &\sum_{t \in \mathcal{T}_c} (y_t - \hat{y}_t)^2 / N_c \\ &= \int_{\theta_c}^1 (y - \theta_c)^2 p(y) dy + \int_{-1}^{-\theta_c} (y - (-\theta_c))^2 p(y) dy, \\ &= 2 \int_{\theta_c}^1 (y - \theta_c)^2 p(y) dy, \\ &= 2b^2 e^{-\frac{\theta_c}{b}} - (b_1^2 + b^2 - \theta_c(2b_1 - \theta_c)) e^{-\frac{1}{b}}. \end{aligned} \quad (5)$$

Substituting Eqs. (4) and (5) with Eq. (3) into Eq. (2), SNR ψ becomes a function of two parameters, b and θ_c , as

*e-mail: Tachioka.Yuki@eb.MitsubishiElectric.co.jp

†e-mail: Narita.Tomohiro@ct.MitsubishiElectric.co.jp

‡e-mail: Ishii.Jun@ab.MitsubishiElectric.co.jp

$$\psi_b(\theta_c) = 10 \log_{10} \frac{((\theta_c + b)^2 + b^2)e^{\frac{1-\theta_c}{b}} - b_1^2 - b^2}{2b^2e^{\frac{1-\theta_c}{b}} - b_1^2 - b^2 + \theta_c(2b_1 - \theta_c)}. \quad (6)$$

Note that b can be derived by the shape fitting of nonclipped speech *a priori*.

4. ASR performance estimation using logistic regression

The ASR performance (word recognition accuracy) A can be estimated, using a logistic regression [2], as

$$A(x) = \frac{\alpha}{1 + \exp(-\beta(x - \gamma))}, \quad (7)$$

where α , β , and γ are parameters of regression and x is an objective measure. In this paper, x is the SNR in Eq. (2) and PESQ, which evaluates speech quality, ranges from 0.5 to 4.5 [4]. Under noisy environments, a previous study [2] revealed that PESQ has a good correlation with the ASR performance. α can be determined from the accuracy, A_o , for original (nonclipped) speech as

$$\alpha = \begin{cases} A_o & (\text{for SNR}), \\ A_o(1 + \exp(-\beta(4.5 - \gamma))) & (\text{for PESQ}), \end{cases} \quad (8)$$

because $A(\infty)$ for SNR and $A(4.5)$ for PESQ must be equal to A_o . β and γ are estimated by the minimum square errors criterion. α represents the difficulty of the task; β is the sensitivity of the performance against x ; γ is the point where recognition accuracy becomes half, i.e., the point of inflection, and indicates the robustness against clipping.

5. Speech recognition experiments on clipped speech

5.1. Setup

We evaluated the JEIDA-JCSD (B-set) dataset (16 kHz sampling) consisting of 100 Japanese city names (e.g., ‘‘Sapporo’’) spoken by 20 male and 20 female speakers. Clipped utterances were artificially made with different θ_c values in the range from 0.1 to 0.9. We prepared two tasks, a large vocabulary (155,592 words) task and a small vocabulary (100 words) task, because the influence of clippings can depend on the difficulty of the task. We used the Julius (ver. 4.2.1) software [6] for decoding. The acoustic model was the 64 mixture phonetic-tied mixture [7] context-dependent

Fig. 1 SNR and PESQ of clipped signals in terms of clipping level θ_c , which is defined in Eq. (1), with SNR estimated using Eq. (6).

triphone HMM attached to Julius. The number of states was 3,131, and the number of Gaussians was 8,256. The acoustic features were 12-dim MFCCs, their Δ , and Δ log power. The total number of dimensions was 25.

5.2. ASR performance estimation

Figure 1 shows the SNR and PESQ for clipped utterances. SNR and PESQ had almost linear relationships with clipping level, and they dropped with decreased clipping level. In the figure, the dotted line shows the SNR estimated using Eq. (6) with $b = 0.15$. The estimated SNR matched the actual SNR.

Figure 2 shows the relationships of (a) SNR and (b) PESQ with word recognition accuracy. Clipping levels θ_c were arranged from 0.1 to 0.9. When θ_c was over 0.7, the recognition accuracy did not decrease significantly and speech quality degradation was inaudible. Otherwise, the recognition accuracy decreased; this tendency was more significant for the large vocabulary task. SNR and PESQ had clear correlations with the word recognition accuracy. The dotted line in the figures is a fitted logistic regression curve estimated by the minimum square errors criterion. Table 1 shows the parameters. A logistic regression can express word recognition accuracies well with a small number of parameters.

Fig. 2 Recognition rate for small vocabulary (100 words) and large vocabulary (155,592 words) tasks in terms of (a) SNR and (b) PESQ. Clipping level θ_c varies from 0.1 to 0.9. For speech recognition performance estimation, logistic regression was used.

Fig. 3 Confusion matrix of Japanese phonemes (a) without clippings and (b) with clippings ($\theta_c = 0.2$), where ‘I’ stands for insertion and ‘D’ stands for deletion. Contour map of correct ratio. The sum along each line is normalized to one.

Table 1 Parameters used in performance estimation.

	SNR		PESQ	
	Large	Small	Large	Small
A_o	63.0	93.0	63.0	93.0
β	0.3		1.7	
γ	0	5.8	1.8	2.8

5.3. Phonemes error tendencies

Figure 3 shows the confusion matrices of 23 Japanese phonemes (a) without clippings and (b) with clippings, respectively. The number of errors increased among vowels. The number of errors where consonants were recognized as vowels was greater than that of errors where vowels were recognized as consonants.

The number of insertion errors increased more than that of deletion errors, especially for consonants. The whitening effect due to clipping caused the voices to become similar to the noise, and the voices did not match the acoustic models. Consonants were more susceptible to these phenomena than vowels, because consonants are originally more similar to the noise than vowels.

6. Conclusion

Speech recognition performance was evaluated by changing the clipping level while keeping the other factors the same. Experiments showed that SNR and PESQ correlate

with the ASR performance of clipped speech as well as noisy speech cases. Moreover, we derived a theoretical estimation formula of SNR after clipping and showed its high accuracy. Confusion matrices revealed that some typical errors were caused by clipping, and consonants were more susceptible to clipping than vowels.

References

- [1] S. Maymon, E. Marcheret and V. Goel, “Restoration of clipped signals with application to speech recognition,” *Proc. Inter-speech*, pp. 3294–3297 (2013).
- [2] T. Yamada, M. Kumakura and N. Kitawaki, “Performance estimation of speech recognition system under noise conditions using objective quality measures and artificial voice,” *IEEE Trans. Audio Speech Lang. Process.*, **14**, 2006–2013 (2006).
- [3] T. Fukumori, M. Nakayama, T. Nishiura and Y. Yamashita, “Estimation of speech recognition performance in noisy and reverberant environments using PESQ score and acoustic parameters,” *Proc. APSIPA ASC*, 2942163 (2013).
- [4] ITU-T Rec P862, “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow band telephone networks and speech codecs,” <http://www.itu.int/rec/T-REC-P862/> (2001).
- [5] S. Gazor and W. Zhang, “Speech probability distribution,” *IEEE Signal Process. Lett.*, **10**, 204–207 (2003).
- [6] A. Lee, T. Kawahara and K. Shikano, “Julius — An open source real-time large vocabulary recognition engine,” *Proc. Euro-speech*, pp. 1691–1694 (2001).
- [7] A. Lee, T. Kawahara, K. Takeda and K. Shikano, “A new phonetic tied-mixture model for efficient decoding,” *Proc. ICASSP*, pp. 1269–1272 (2000).