Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 72 (2015) 218–231

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Detection and reconstruction of clipped speech for speaker recognition

Fanhu Bie, Dong Wang, Jun Wang, Thomas Fang Zheng *

Center for Speech and Language Technologies, Division of Technical Innovation and Development, Tsinghua National Laboratory for Information Science and Technology, China
Center for Speech and Language Technologies, Research Institute of Information Technology, Tsinghua University, China
Department of Computer Science and Technology, Tsinghua University, China

Received 2 February 2015; received in revised form 19 June 2015; accepted 22 June 2015
Available online 2 July 2015

Abstract

Clipping is often observed in speech acquisition, due to the limited numerical range or the non-linear compensation of recording devices. The clipping inevitably changes the spectrum of speech signals, and thus partially distorts the speaker information contained in the signal. This paper investigates the impact of signal clipping on speaker recognition, and proposes a simple yet effective clipping detection approach as well as a signal reconstruction approach based on deep neural networks (DNNs). The experiments are conducted on the core test of the NIST SRE2008 task by simulating clipped speech at various clipping rates. The results show that clipping does impact the performance of speaker recognition, but the impact is rather marginal unless the clipping rate is larger than 80%. We also find that the simple distribution-based detection method is capable of detecting clipped speech with a higher accuracy, and the DNN-based reconstruction can achieve promising performance gains for speaker recognition on clipped speech.

Keywords: Speech clipping; GMM–UBM; i-vector; DNN; Speaker recognition

1. Introduction

After decades of research, current speaker recognition has achieved rather satisfactory performance, given that the enrollment and test utterances are sufficiently long and the quality is sufficiently high (Campbell et al., 2006; Bimbot et al., 2004). However, when the signals are corrupted, the performance of a speaker recognition system will generally degrade significantly.

A lot of research has been conducted to improve the robustness of speaker recognition, for example in conditions with mismatched channels and strong noises. Various feature-based approaches (such as feature adaptation) or model-based approaches (such as channel synthesis or channel factorization) have been demonstrated effective to mitigate the impact of signal corruptions. For a particular corruption, signal clipping, however, the research is still very limited. Denoting the maximum amplitude of a signal by E_m, and the maximum sampling value of the recording device by E_q, signal clipping is observed when E_m exceeds E_q, resulting in the received sample ceiled at E_q. In some circumstances, the recording device adjusts the recording gain automatically when high-volume input is detected. In this case, the received sample may be ceiled at a value E_c that is lower than E_q. We define E_c as the ‘clipping value’ in this paper. Fig. 1 illustrates the clipping phenomenon of a sine signal whose sample size is 8 bits, and Fig. 2 shows two real-world clipped speech signals with and without automatic gain adjustment, respectively.

Although often ignored in speaker recognition, the clipping phenomenon has gained much attention in other fields

* Corresponding author at: Center for Speech and Language Technologies, Division of Technical Innovation and Development, Tsinghua National Laboratory for Information Science and Technology, China.
E-mail address: fzheng@tsinghua.edu.cn (T.F. Zheng).

http://dx.doi.org/10.1016/j.specom.2015.06.008
0167-6393/© 2015 Elsevier B.V. All rights reserved.

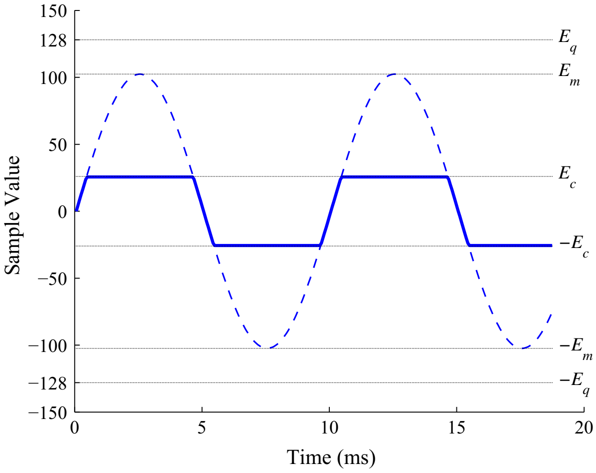

Fig. 1. Speech clipping of a sine signal with 8-bit precision.

of speech processing. For example, Kates et al. (1994) conducted a systematic study on the impact of signal clipping on speech quality. Licklider (1946) reported that clipped speech could be perfectly intelligible, even if the clipping value E_c is 10% of the amplitude of the original signal, although the speech quality reduction can be noticed. Crain and Van Tasell (1994) found that the clipping value at which the intelligibility of speech starts to be significantly affected coincides with the clipping value at which the quality of the speech is judged to be unacceptable. Kitic et al. (2013) and Harvilla and Stern (2014) presented a detailed analysis on properties of clipped speech and its impact on automatic speech recognition (ASR) and found that clipping might cause noticeable signal distortion that should be carefully compensated for. A similar study was also conducted in Kitic et al. (2013). Recently, Tachioka found that the impact of signal clipping on human perception (in terms of perceptual evaluation of speech quality (PESQ)) and ASR performance (in terms of word accuracy) were closely related, and the latter could be well predicted from the former by logistic regression (Tachioka et al., 2014).

In order to mitigate the impact of clipping, researchers have proposed some approaches to detecting and/or reconstructing the original signal, particularly in the ASR community (Rabiner, 1989). To clipping detection, Deng et al. (2013) proposed an approach based on kernel Fisher linear discriminant analysis, and Eaton and Naylor (2013, 2014) studied a perceptual codec for it. Aleinik and Matveev (2014) proposed a method based on histograms of signal values in the time domain. This approach was quite similar to the method proposed in this paper (refer to Section 4) and had been demonstrated rather simple and effective. For clipping reconstruction, a straightforward solution is to employ a regression model to predict the original values of clipped samples, for instance, the linear predictive coding method (Bradbury, 2000). Janssen et al. (1986) used the EM algorithm to perform the reconstruction with an iterative procedure, where the criterion was to minimize the residual errors. Similarly, Selesnick (2013) proposed a de-clipping approach based on the principle of minimizing the third derivative of the reconstructed signal. Kitic et al. (2013) proposed a reconstruction approach based on sparse analysis. A similar approach was proposed in Adler et al. (2012), where distortion was separated and eliminated by sparse decomposition using the orthogonal matching pursuit (OMP) algorithm. This approach was shown to be effective for various distortions, including clipping, impulse noises and pack loss. Other related approaches involve sample interpolation (Dahimene et al., 2008; Lagrange et al., 2005), bandwidth extension (Dietz et al., 2002; Smaragdis et al., 2009; Moussallam et al., 2010), and concealment (Perkins et al., 1998; Ofir et al., 2007). Note that almost all the above-mentioned reconstruction methods are based on linear models, whereas the distortion caused by clipping is obviously nonlinear. A better de-clipping approach is desired, preferably nonlinear.

This paper studies the impact of clipped speech on speaker recognition. From the results obtained in the ASR research as mentioned above, one can conjecture that clipping should impact speaker recognition if it is aggressive. However, speaker recognition and ASR are two fundamentally different tasks, and it is interesting to investigate how the clipping impacts speaker recognition. In addition, encouraged by the performance gains obtained in ASR with clipped speech reconstruction, this paper

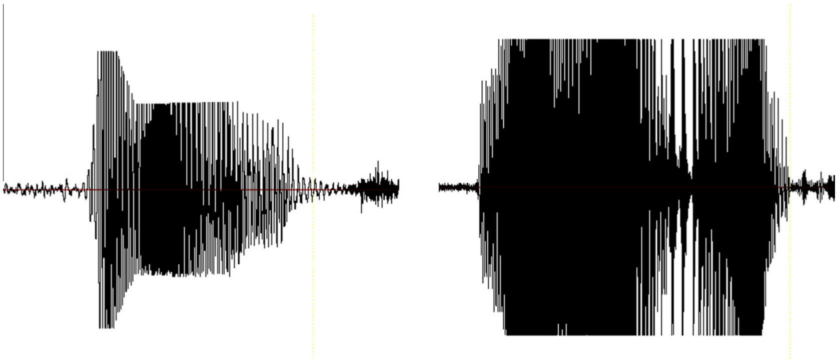

Fig. 2. Speech clipping with (left) or without (right) automatic gain adjustment.