RESEARCH

Detection and Reconstruction of Clipped Speech in Speaker Recognition

Fanhu Bie[??]
, Dong Wang
, Jun Wang
and Thomas Fang Zheng

[??] Correspondence:
biefh@cslt.riit.tsinghua.edu.cn
Center for Speech and Language Technologies, Division of Technical Innovation and Development, Tsinghua National Laboratory for Information Science and Technology;
Center for Speech and Language Technologies,
Research Institute of Information Technology, Tsinghua University;
Department of Computer Science and Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China
Full list of author information is available at the end of the article

Abstract

Signal 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 signal, 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) to reconstruct the signal from a clipped one. 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 is highly aggressive with the clipping rate 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 (also known as voiceprint recognition) has achieved rather satisfactory performance, given that the enrollment and test utterances are sufficiently long and the quality is sufficiently high, that is to say, the speech signals are well recorded and the noise corruption is limited [1, 2]. 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 loud 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 impact of some corruptions such as channel mismatch and background noises. 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 facility 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 facility will adjust 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 sampled at $8-bit$ precision, and Fig. 2 shows two real-world clipped speech signals with and without automatic gain adjustment, respectively.

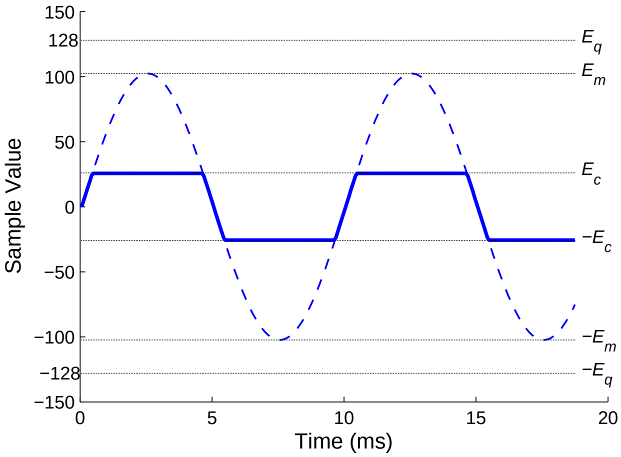

Figure 1 Speech clipping of a sine signal with 8-bit precision.

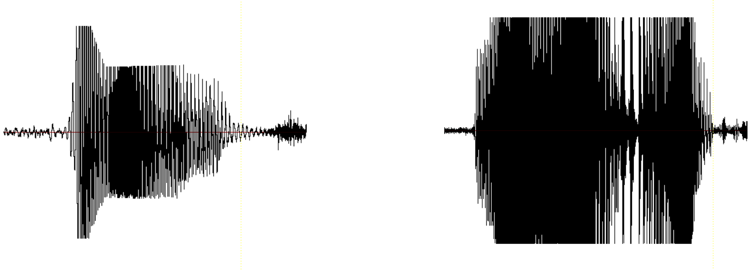

Figure 2 Speech clipping with (left) or without (right) automatic gain adjustment.

Although often ignored in speaker recognition, the clipping phenomenon has gained much attention in other fields of speech processing. For example, [3] conducted a systematic study on the impact of signal clipping on speech quality. [4] reported that clipped speech could be perfectly intelligible, even if the clipping value E_c was 10% of the amplitude of the original signal, though the speech quality reduction could be noticed. [5] found that the clipping value at which the intelligibility of speech started to be significantly affected coincides with the clipping value at which the quality of the speech was judged to be unacceptable. [6] and [7] 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 before the speech is fed to the ASR system. A similar study was also conducted in [6].

In order to mitigate the impact of clipping, researchers have proposed some approaches to reconstructing the original signal, particularly in the ASR community [8]. A straightforward solution was to employ a regression model to predict the original values of clipped samples, for instance, the linear predictive coding method [9]. [10] used the EM algorithm to perform the reconstruction with an iterative procedure, where the criterion was to minimize the sum of squares of the residual errors. Similarly, Selesnick [11] proposed a de-clipping approach based on the principle of minimizing the third derivative of the reconstructed signals. [6] proposed a reconstruction approach based on sparse analysis. A similar approach was proposed in [12], where distortion was separated and eliminated by sparse decomposition using the orthogonal matching pursuit (OMP) algorithm. This approach was effective for various distortions, including clipping, impulse noises and pack loss. Other related approaches involved sample interpolation [13, 14], bandwidth extension [15, 16, 17], and concealment [18, 19]. Note that almost all the above-mentioned reconstruction methods were based on linear models, whereas the distortion caused by clipping is obviously nonlinear. A better de-clipping approach preferably nonlinear, is desired.

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 proposes a novel clipping reconstruction method based on deep neural networks (DNNs), which can learn the complex nonlinear distortion associated with clipping, and therefore is highly powerful for recovering clipped speech.

The experiments were conducted on the core test of the NIST SRE2008 task, by simulating clipped speech at various clipping values. Two speaker recognition systems were constructed, based on the conventional Gaussian mixture model-universal background model (GMM-UBM) architecture and the state-of-the-art i-vector model, respectively. The results show that clipping does impact speaker recognition, but the impact is rather marginal unless the clipping is aggressive. Specifically, we observe that the recognition performance largely remains if the clipping value is higher than 20% of the amplitude of the original signal ($E_c \geq 0.2E_m$). In addition, the i-vector system is clearly more robust against clipping. We also see that a simple distribution-based detection approach is capable of detecting clipping at a high accuracy, and the DNN-based clipping reconstruction can offer promising performance gains for speaker recognition.

The rest of the paper is organized as follows: Sec. 2 analyzes the spectrum distortion caused by speech clipping in a basis of a sine signal. Sec. 3 discusses the impact of clipped speech on two speaker recognition systems. A distribution-based clipping detection approach is proposed in Sec. 4 and the DNN-based clipping reconstruction approach is proposed in Sec. 5. The whole paper is concluded by Sec. 6.

2 Analysis on clipping

In order to study the clipping phenomenon, we first simulate some clipped speech and study the spectrum distortions it caused. The simulation is helpful since the clipping rate can be controlled in such a way that the distortion and its impact on human auditory systems can be studied systematically.

2.1 Definitions

For simplification, we first define some quantities that will be used in this study. As mentioned in Sec. 1, we denote the concept of the upper limit of the sample range as E_q, define the maximum amplitude of a speech signal as E_m and that of the clipped speech signal as E_c, which is also mentioned as the clipping value in this paper. The 'clipping rate' γ is defined as the proportion of the 'clipped amplitude', given by

$$\gamma = 1 - \frac{E_c}{E_m}.$$

.

The bigger the γ is, the more proportion of the 'amplitude' is cut off. Specially, $\gamma = 0$ means that no portion of the signal is clipped.

Note that in practice, clipping occurs when the sample value exceeds the sample range and so E_c is the upper limit of the sample value E_q, and so the maximum amplitude of the original signal E_m is unknown. The definitions of E_m and E_c are just for simulation, and in this case, both E_c and E_m are known and smaller than E_q.

2.2 Analysis on a clipped sine signal

We start the study on clipping by observing the spectrum distortion caused by clipping on a simple sine signal. The reason of choosing the sine signal is two-fold: firstly, it is the simplest signal with only one frequency component and hence is easy to study; secondly and more importantly, any speech signal can be decomposed into a weighted sum of sine signals of different frequencies according to the Fourier transform, so the study on sine signals will shed lights on the impact of clipping for any natural speech signals.

In this study, the frequency of the sine signal is selected to be $100Hz$, and the amplitude is fixed to 1, written mathematically as

$$s(t) = sin(\frac{2\pi}{100}t). \tag{1}$$

This signal is clipped with a variety of clipping rates. Given a clipping rate γ, the clipped signal of $s(t)$ is given as follows:

$$\tilde{s}(t) = \begin{cases} s(t) & (|s(t)| < 1 - \gamma) \\ sgn(s(t)) \times \alpha & (|s(t)| \geq 1 - \gamma) \end{cases}, \tag{2}$$

where $sgn(\cdot)$ is the sign function.

The time and frequency domain representations of the clipped signals are shown in Fig. 3. We can see that clipping does not change the fundamental frequency,

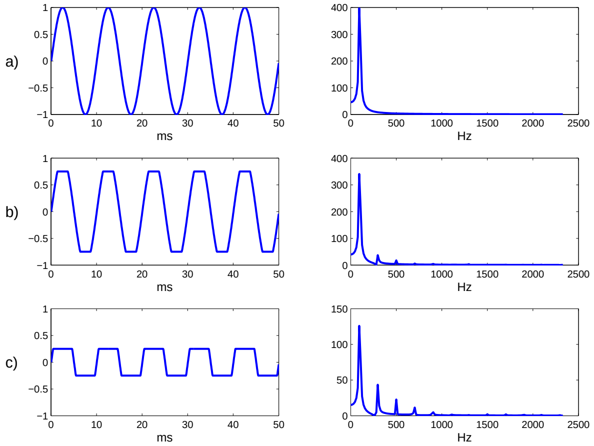

Figure 3 The waveform (left) and spectrum (right) of the sine signal $s(t)$ as a function of clipping rate. a) Original sine signal with $\gamma = 0$; b) Clipped signal with $\gamma = 0.25$; c) Clipped signal with $\gamma = 0.75$.

however it does introduce extra harmonics that attenuate the energy at the original (fundamental) frequency. Obviously, a larger clipping rate γ leads to a more aggressive clipping, and more spectrum distortion.

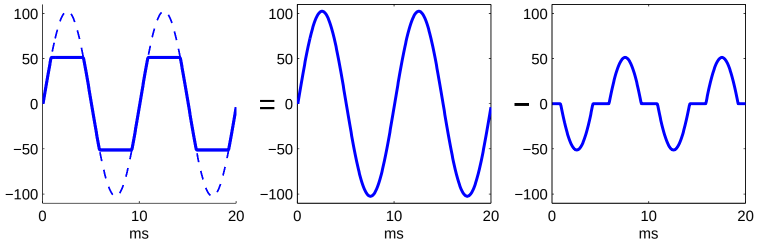

Figure 4 Illustration of decomposition of a clipped sine signal with $\gamma = 0.5$.

Let's derive these properties in a more rigorous way. First note that a clipped sine signal can be decomposed into the original sine signal and a symmetric periodic impulse signal, as shown in Fig. 4. Mathematically, the decomposition can be formulated as

$$\tilde{s}(t) = sin(\omega t) - g(t), \tag{3}$$

where ω is the angel frequency, $g(t)$ is the symmetric periodic impulse signal, which can be further decomposed into two periodic impulse signals as

$$g(t) = f(t - T/4) - f(t + T/4),$$

where $f(t)$ is the periodic impulse signal whose period is T. Note that T, the period of the impulse signal, is determined by the 'host signal' $sin(\omega t)$, where $T = \frac{2\pi}{\omega}$. In one period, the periodic rectangular function can be written as

$$f(t) = \begin{cases} sin(\omega t + \pi/2) - sin(\omega\tau/2 + \pi/2) & |t| \leq \tau/2 \\ 0 & \tau/2 < |t| \leq T/2 \end{cases},$$

where τ is the width of the impulse signal. The decomposition process is illustrated in Fig. 5.

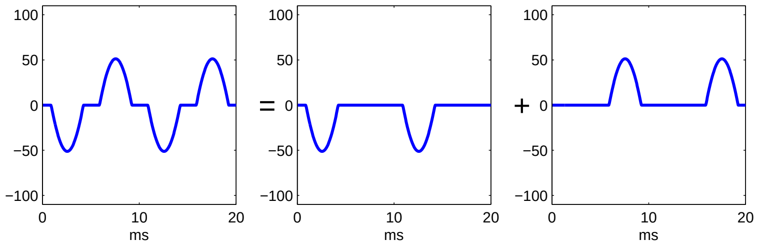

Figure 5 Illustration of decomposition of a symmetric periodic impulse signal with $\gamma = 0.5$.

Let $\{a_n\}$ represent the Fourier series of $f(t)$,

$$f(t) = \sum_{n=-\infty}^{\infty} a_n e^{jn\omega t},$$

and the symmetric impulse signal can be written by

$$\begin{aligned} g(t) &= \sum_{n=-\infty}^{\infty} a_n(e^{jn\omega(t-T/4)} - e^{jn\omega(t+T/4)}) \\ &= \sum_{n=-\infty}^{\infty} (-2j)a_n sin(n\omega T/4)e^{jn\omega t}. \end{aligned}$$

Combing with Equ. 3, we reach the main result for a clipped sine signal as

$$\tilde{s}(t) = sin(\omega t) + \sum_{n=-\infty}^{\infty} 2a_n sin(\frac{n\pi}{2})e^{j(n\omega t + \frac{\pi}{2})}. \tag{4}$$

From Equ. 4, the following can be derived: (1) signal clipping introduces and only introduces harmonic frequencies of the original frequency; (2) the harmonics appear at frequencies corresponding to odd n; (3) the spectrum distortion is determined by $\{a_n\}$, which is in turn determined by the clipping width τ. Obviously, these derivations are consistent with the observations in Fig. 3.

For an actual speech signal, analyzing the impact of clipping is much more complex. Particularly, different frequency components in a real signal possess different values of amplitude and phase, leading to very different $\{a_n\}$ in Equ. 4 and resulting in significant inter-frequency interference. A complex non-linear model is probably required if one intends to reconstruct the original signal from a clipped speech segment.

2.3 human perception evaluation

Although the impact of clipping in spectrum domain is observed, we are also interested in the impact of clipping on human perception. In this study, we select 10 speech signals and use different clipping rates to generate clipped speech. The speakers are asked to listen to the utterances clipped at different clipping rates (including the case when $\gamma = 0$) and tell whether they are original or clipped. If a clipped speech is recognized as the original speech, a *non-difference* is counted. By calculating the average *non-difference* rate, which is the percentage of the *non-difference* in all the listening trials, one can tell whether or not a particular clipping rate causes significant impact on human perception.

There are 5 speakers are involved in these experiments, and the non-difference rates with various clipping rates are shown in Fig. 6, where the horizontal axis is the clipping rate where the vertical axis is the *non-difference* rate. It can be seen from the figure that when the clipping rate is relatively small, clipping produces no significant impact on human perception. However, when the clipping rate is larger than 80%, one can perceive clear difference between a clipped speech and the original one. This result is consistent with the findings in [4].

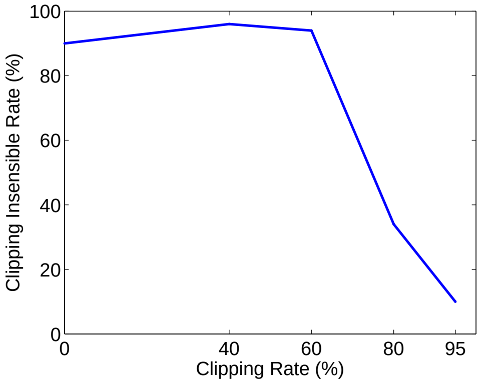

Figure 6 Percentage of clipping insensible listeners as a function of clipping rate in a perception evaluation.

3 Impact of clipping on speaker recognition

This section studies the impact of speech clipping on the performance of speaker recognition. The recognition experiments are conducted and evaluated on the data of the NIST2008 core test, yet focusing on the same channel condition only. It is a verification task, though we assume the conclusions obtained with verification are largely shared with identification. The speech data are interviews recorded by microphones, at an $8kHz$ sampling rate with 16-bit precision. The data set involves 171 female speakers and 996 test utterances in total. The enrollment and test utterances each lasts about 3 minutes.

Two speaker recognition systems are constructed: one is a traditional GMM-UBM system while the other one is an i-vector system. The Fisher database is used as the development set to train the UBM for both systems, and the loading matrix T for the i-vector system. The PLDA model [20, 21] is also used to further improve the performance of the i-vector system, and the model is trained using the same development data.

The acoustic features are 60-dimensional MFCCs, including 19 dimensional M-FCCs and 1 dimension of energy, plus their first- and second- order derivatives. The UBM/GMM contains $2,048$ Gaussian components, and the dimension of i-vectors is 400.

3.1 Speaker recognition results

We report speaker recognition results with the GMM-UBM system and the i-vector system. The enrollment and test utterances are clipped at four clipping rates (40%, 60%, 80%, and 90%), and then speaker recognition is conducted with the clipped speech at various clipping rates, with performance evaluated in terms of the equal error rate (EER).

The EER results of the GMM-UBM system are reported in Fig. 7, and the results of the i-vector system are reported in Fig. 8. Each group in the figures represents a particular clipping condition on the enrollment utterances, and each bar reports the result of a particular clipping condition on the test utterances.

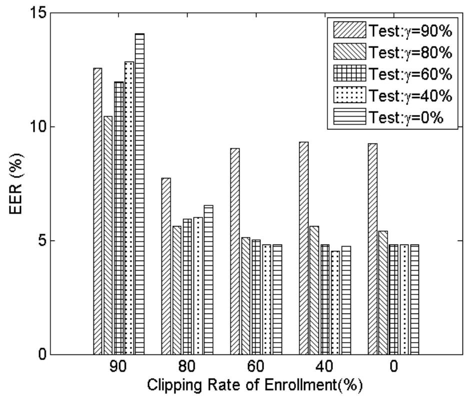

Figure 7 Results in EER for the GMM-UBM systems, with enrollment and test utterances clipped at 4 different clipping rates.

From the results, we can see that speech clipping indeed impacts performance of speaker recognition, for both the GMM-UBM system and the i-vector system. However, the impact is rather marginal unless the clipping rate is larger than 80%. This result is consistent with the observation of the subjective listening test reported in Sec. 2.3, and suggests that the degradation caused by clipping is not as significant as one may imagine at the first glance.

When comparing the GMM-UBM system with the i-vector system, it can be seen that the i-vector system outperforms the GMM-UBM system in all the test

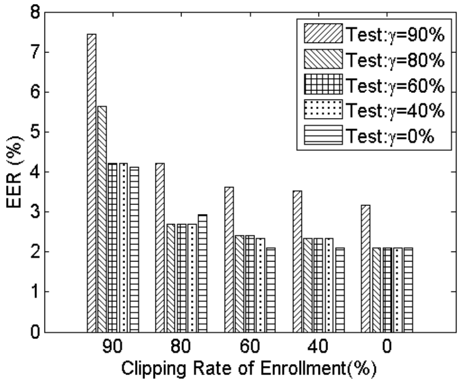

Figure 8 Results in EER for the i-vector system, with enrollment and test utterances clipped at 4 different clipping rates.

conditions, and it exhibits more robustness against clipped speech. For example, keeping the enrollment utterances unclipped (γ=0) and compare the performance of the two systems when the clipping rate of the test utterances is set to 80%, we observe clear performance degradation with the GMM-UBM system, however for the i-vector system the result is almost equal to the one with the unclipped test utterances. If the clipping rate goes up to 90%, the performance of the GMM-UBM system decreases by 92%, while the i-vector system decreases by 50% only. In addition, when the clipping is highly aggressive with the enrollment utterances, serious performance degradation is observed with the GMM-UBM system, no matter how big the clipping rate is for the test utterances. For the i-vector system, however, the performance does not significantly degrade. The advantage of i-vector systems on clipped speech may be contributed to the low dimensionality of the i-vector space which retains information that is mostly related to speaker characteristics. Furthermore, the PLDA model may contribute to the discriminative power of i-vectors and lead to a model more robust against clipping.

3.2 Visualization for speaker model change

To gain further insight on how speech clipping impacts speaker recognition, this section visualizes the change on speaker models caused by clipping. Two visualization approaches are employed: dimension selection and dimension reduction. In the first approach, two dimensions are randomly selected from the parameter supervector of a speaker model to represent the speaker, and in the second approach, the t-SNE algorithm [22, 23, 24, 25] is used to project the entire speaker model to a two-dimensional space.

3.2.1 Visualization by dimensions selection

We first investigate the dimension selection approach, where two dimensions are selected from the supervector of a speaker model to represent the speaker, which means the i-vector itself in the i-vector model, or the concatenation of the mean vectors of the components in the GMM-UBM model. Fig. 9 shows the result of the i-vector system with different clipping rates. Each speaker is represented by the first two dimensions of its enrollment i-vector, and the model change caused by speech clipping can be read from the displacement between the two-dimensional i-vectors enrolled with and without clipping.

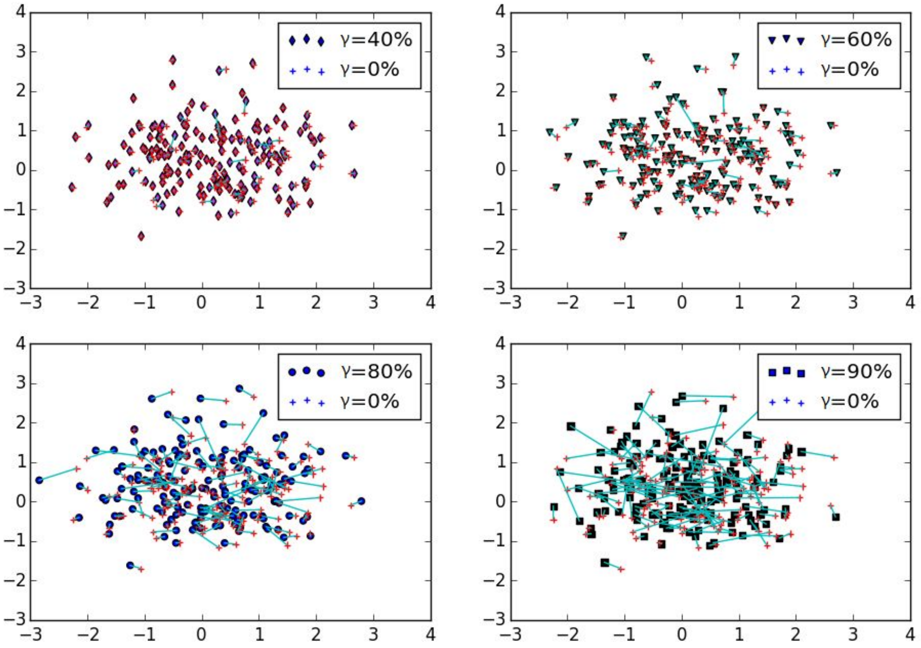

Figure 9 Model change in the i-vector system with different clipping rates. The first two dimensions of a speaker's i-vector are selected to represent the speaker.

From Fig. 9, it can be see that a more aggressive clipping leads to a more significant displacement in the i-vector space. In the case of $\gamma \geq 60\%$, the impact of clipping is so significant that the speakers are difficult to be distinguished. The impact of speech clipping can be also measured by an i-vector distortion rate. Let $D(\gamma)$ denote the sum of the i-vector displacements at a particular clipping rate γ, and D_v as the summed length of the original i-vectors (enrolled with unclipped speech). The i-vector distortion rate is defined as

$$\Delta(\gamma) = \frac{D(\gamma)}{D_v}. \tag{5}$$

Fig. 10 presents the curve of $\Delta(\gamma)$. It can be seen that when the clipping rate is less than 60%, the distortion is relative small, and when the clipping rate is larger than 60%, the distortion increases sharply. This is consistent with the results of the human perception in Sec. 2.3 and the speaker recognition results reported in Sec. 3.1.

For the GMM-UBM system, each speaker is represented by a GMM. To visualize a speaker, we randomly select a Gaussian component from the speaker's GMM and

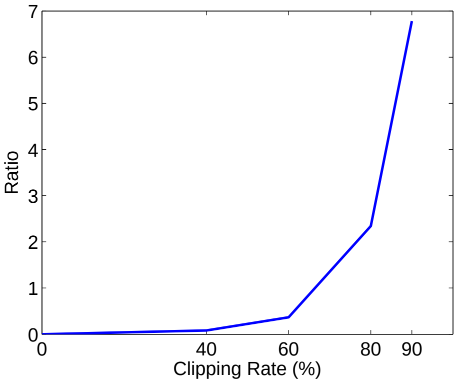

Figure 10 I-vector distortion rate $\Delta(\gamma)$ as a function of clipping rate.

then select two dimensions of its mean vector as a two-dimensional representation of the speaker (speaker vector). Again, the impact of speech clipping is represented by the displacement in the speaker vector space. The results are shown in Fig. 11. Similar to the case of the i-vector system, we observe that when the clipping rate is relatively low (less than 60%), the speaker models are not much impacted; however, when the clipping rate is larger than 60%, the speaker models are changed such significantly that speakers are largely undistinguishable.

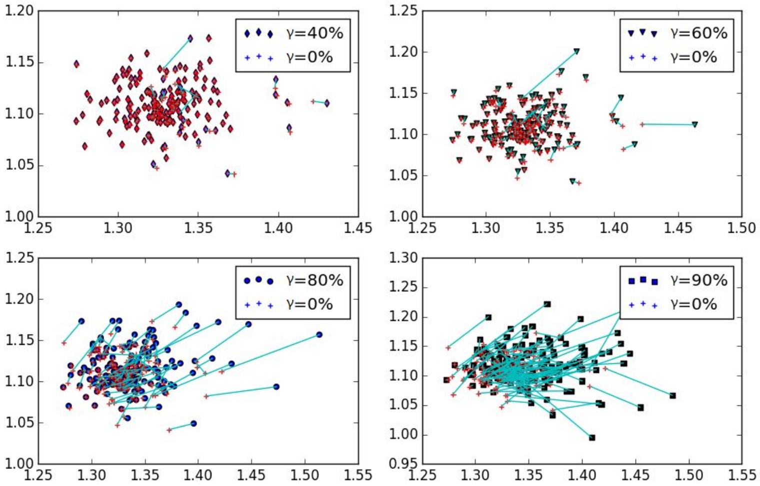

Figure 11 Model change in the GMM-UBM system with different clipping rates. A Gaussian component is randomly selected, and the first two dimensions are selected from the mean vector of the selected Gaussian to represent speakers.

3.2.2 Visualization by t-SNE

The dimension selection process selects two dimensions randomly. To have a more confident visualization, we employ the t-SNE technique ito project speaker models to a two-dimension space. This approach can be regarded as a dimension reduction method, with which the affinity property of speaker models in the high-dimension space is reserved as much as possible when projecting to a two-dimensional space [22, 23, 24, 25].

For the i-vector system, visualization with t-SNE is straightforward: we simply project the 400-dimensional enrollment i-vectors into two-dimensional vectors and observe the displacement caused by speech clipping, as in the previous section. Fig. 12 presents the results, where i-vectors enrolled with different clipping rates are drawn together. Again, a clear impact on speaker models can be observed if clipping rate is large than 60%.

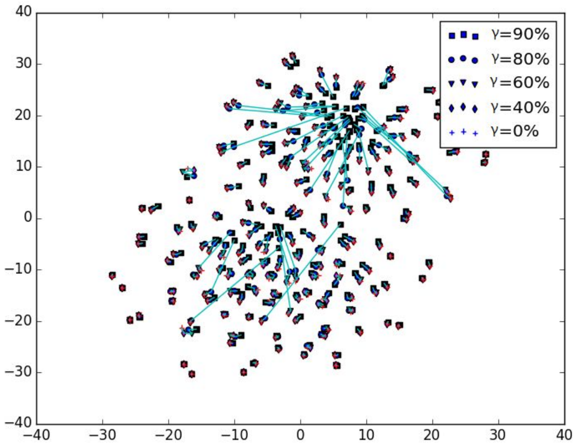

Figure 12 Model change in the i-vector system. I-vectors are projected into two-dimensional vectors by using t-SNE.

For the GMM-UBM system, a speaker is represented by the supervector of its GMM model. Due to the huge dimensions, it is impossible to run t-SNE on supervectors directly. We therefore choose the first two dimensions of the mean vector of each GMM component and concatenate them together to form a partial supervector, and run t-SNE with this representation[1]. The results are shown in Fig. 13. It can be seen that the GMM-based system is more sensitive to signal clipping than the i-vector system. For both the i-vector and the GMM-UBM systems, it seems that that if the clipping rate is small, the change on models caused by clipping is relative small and does not impact the discrimination among speakers. This explains the results obtained in Sec. 3.1 from another view.

4 Clipping detection

Intuitively, a number of approaches can be employed to improve speaker recognition for clipped speech data. Firstly, the clipped speech should be used in enrollment, so that part of the patterns of clipped speech can be learned. We have seen in

[1] We also run the experiment with different choice of the mean vector dimensions, and similar results have been obtained.

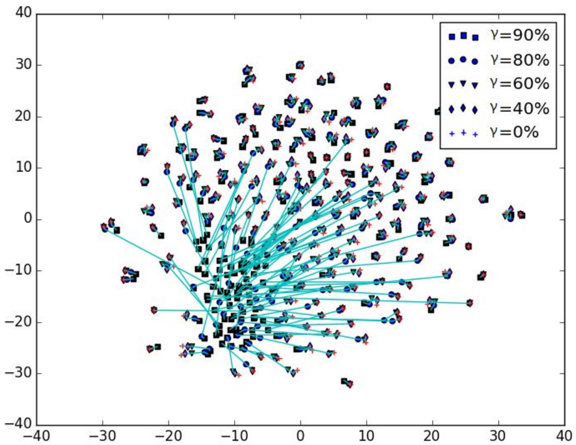

Figure 13 Model change in GMM-UBM system. Speaker GMMs are projected into two-dimensional vectors by using t-SNE.

Sec. 3 that performance with enrollment and test data at the same clipping rate is indeed slightly superior to the cases where the clipping rates are highly mismatched. Secondly, by learning a transform that maps clipped speech to the original speech, the original speech can be reconstructed. This approach is highly promising and we will discuss it in the next section. Thirdly, we can detect the clipped speech and take appropriate actions if clipping is detected, for example, adjust the recording facilities or re-record the speech. This section focuses on clipping detection.

4.1 Clipping detection approach

We propose a simple clipping detection approach based on the time-domain properties of clipped speech. The basic idea is that, when clipping occurs, the distribution of sample values of the signal exhibits different patterns from the ones of a normal unclipped signal. To verify this difference, we select an unclipped speech segment and a speech segment clipped at a clipping rate of 60%. First divide the sample value range into a number of non-overlapped sub-intervals, and then distribute the signal samples to the sub-intervals according to their sample values. This leads to a distribution of the samples over the sub-intervals. The distribution of the unclipped speech and clipped speech are shown in Fig. 14 and Fig. 15, respectively (the small-valued samples are discarded to avoid the impact of white noise).

Figures show that the distributions of the two types of speech segments are quite different: the majority of samples of the unclipped speech concentrate within the sub-intervals corresponding to middle values, while the distribution of the clipped speech concentrates in the last sub-interval. This difference can be adopted to detect clipped speech segments.

Accordingly, we propose the following clipping detection algorithm:

- Check the maximum amplitude value s_max of all the samples in the given speech signal s, and divide the range $[0, s_max]$ into a predefined number of, say 20 as an example in this paper, sub-intervals $\{\beta_i : i = 1, 2, .., 20\}$.

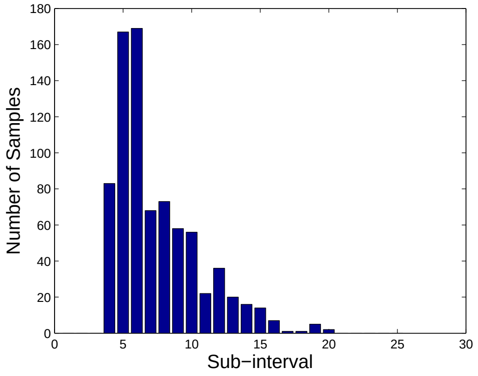

Figure 14 Distribution of unclipped speech segment

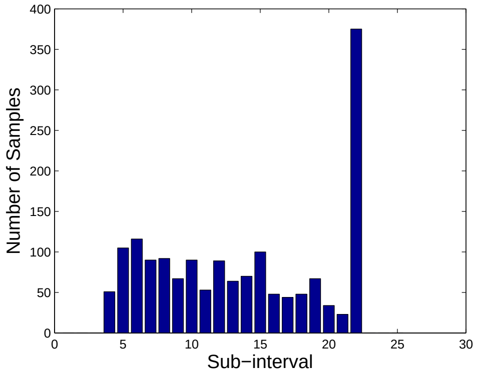

Figure 15 Distribution of clipped speech segment

- Divide the samples of the speech signal s into non-overlapped segments $s = [c_1, c_2, ..., c_K]$ where K is the total segments number of the utterance. Each segment is 0.5-second long and regarded as the unit of detection for the clipping event.
- Given a specific segment c_k, distribute its samples to the sub-intervals $\{\beta_i\}$, according to the sample values.
- Calculate the distribution over the sub-intervals $\{\beta_i\}$, and detect clipped segments by the distribution mass of the last sub-interval. If $\beta_{20} > \epsilon$ where ϵ is a predefined threshold, the segment is regarded as *clipped*, otherwise *normal*.

The performance of the proposed clipping detection is evaluated in terms of the false acceptance rate (FAR) and the false rejection rate (FRR), where FAR reflects the probability that a normal segment is recognized as clipped, and FRR reflects the probability that a clipped segment is recognized as normal. By varying the threshold ϵ, a tradeoff between the FAR and FRR is obtained, leading to a detection error tradeoff (DET) curve.

We randomly select 1-hour speech from the development set to conduct the clipping detection experiments. The DET curves with different clipping rates are shown in Fig. 16. From the figure, it can be seen that the proposed clipping detection method is rather accurate (the EER is less than 1%). Even if the clipping rate is relatively large (40%), a rather high accuracy is still achieved.

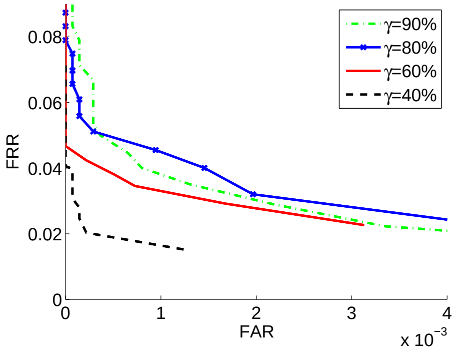

Figure 16 DET curves of the proposed clipping detection.

4.2 Sample removal after clipping detection

Intuitively, the removal of clipped speech segments can prevent the negative impact caused by clipping, and thus may improve the performance of speaker recognition. However, removing clipped segments will also reduce the amount of effective data for either enrollment or testing, and thus perhaps lead to performance degradation. The exact consequence of clipped speech removal is therefore largely unknown and may be dependent on multiple factors such as data profile, clipping rate, and accuracy of the clipping detection. This section presents a pragmatic study.

The experiments are conducted with the same configurations as in Sec. 3, except that the clipped speech segments are discarded in enrollment and testing. Since the i-vector system performs significantly better than the GMM-UBM system, we only report the results for the i-vector system. Fig. 17 presents the EER results based on manual clipped speech transcription, and Fig. 18 reports the EER results based on the label information produced by the clipping detection approach presented in the previous section.

Results show no significant difference between the manual transcription and the automatically generated labels, demonstrating that the proposed clipping detection approach is effective. However, compared with the results in Fig. 8, after clipped speech removal the performance of the i-vector system degrades.

The performance degradation with clipped segment removal might be associated with two factors. Firstly, removing clipped segments reduces the amount of data used for enrollment and testing. Statistics show that the average length of the original enrollment/test utterances is about 3 minutes while after clipped segments removal, the average length is significantly reduced, for example, to 80s with clipping rate at 0.8. Fig. 19 presents the average length of the enrollment data after clipped segment removal. Secondly, the clipped segments are mostly vowels that are assumed to contain richer speaker information, so removing these segments tends to lose the most discriminative part of the data.

As a summary, comparing with the distortion caused by clipping, the information contained in the clipped speech is more important. This suggests that it be better

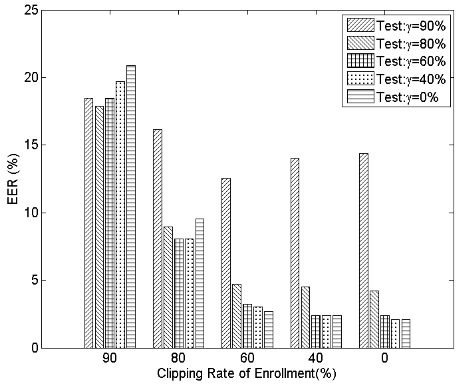

Figure 17 Results in EER based on clipping transcription.

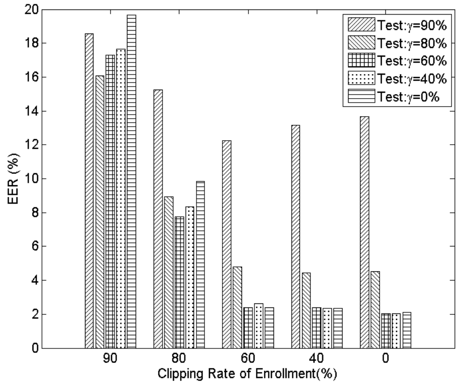

Figure 18 Results in EER based on clipping detection.

to utilize the clipped segments in some way instead of simply discarding them in speaker recognition.

5 Clipped speech reconstruction

According to the previous section, clipped speech segments involve valuable but distorted information for speaker recognition. We therefore try to seek an approach to recovering speaker information from clipped speech signals.

An intuitive approach is to estimate the original signal in time domain. However this turns out to be rather difficult, as speech samples in time domain are highly dynamic. A reasonable approach is to conduct the recovery in feature domain, which might be stable and easy to be integrated in the frontend pipeline of speaker recognition systems. The simplest model is a linear transform, such as constrained maximum likelihood linear regression (MLLR), which maps feature vectors of clipped speech to those of the original speech. This approach, however, is limited by its nature of linearity. As we have discussed, clipping on real speech signals is highly complex and nonlinear, particularly if the clipping rate is unknown or varies.

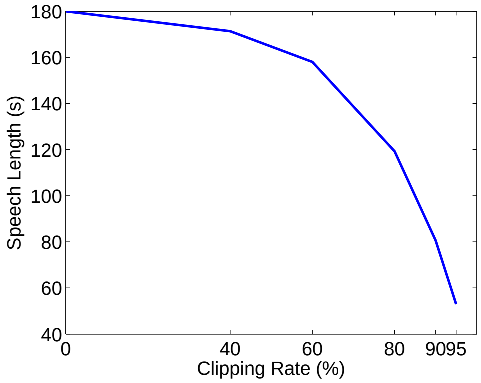

Figure 19 Average length of enrollment utterances as a function of clipping rate.

The DNN model is suitable for learning complex patterns due to its hierarchical architecture and multi-layer non-linearity. In this paper, we propose to use DNNs to recover clipped speech in feature domain.

We note that the impact of speech clipping can be also compensated for in model domain, according to the analysis in Sec. 3.2, by either a linear or a nonlinear model. We will do further research on the model-based compensation in future.

5.1 DNN

Deep neural networks have gained much success in many research fields including speech recognition, computer vision, and natural language processing [26]. A DNN is a neural network (NN) that involves more than one hidden layers. NNs have been studied in the speech community for decades. For example in speech recognition, the NN has been used to substitute the GMM to produce frame likelihood [27], or to produce long-context features that are used to substitute or augment to short-term features, such as MFCCs [28]. Although promising, the NN-based approach did not deliver overwhelming superiority over the conventional approaches. The revolution took place in the ASR community in 2010, after a close collaboration among academic and industrial research groups, including the University of Toronto, Microsoft, and IBM [26, 29, 30]. These researches found that very significant performance improvements can be accomplished with DNNs when appropriate initialization was conducted, for example, by pre-training with restricted Boltzmann machines (RBMs).

Encouraged by the success in ASR, the DNN (and the unsupervised version, deep Boltzmann machine (DBM)) model has been investigated in a wide range of research fields of speech processing, including speech synthesis [31, 32], music pattern analysis [33, 34], speech enhancement [35, 36], voice activity detection [37] and music recommendation [38]. Particularly, a very recent study applies DNN to speaker recognition [39, 40]. The basic idea is to use a DNN model trained for speech recognition to substitute the UBM, so that rich information in phones can be employed to build a more accurate conditional probability model than the GMMs that are trained in an unsupervised way. In this paper, we use the DNN model to recover clipped speech.

5.2 DNN-based clipped speech reconstruction

We construct a DNN that maps features of clipped speech to features of the original speech. We randomly select 50-hour speech data as the development set to train the DNN model. The speech signals of the development set are clipped at various clipping rates, and are fed into the DNN as the input, with the corresponding original speech signal as the output. Specifically, the MFCC features are extracted for both the original speech and the clipped speech, and the features of the clipped speech are fed into the DNN input units frame by frame, with the output units being the corresponding frame of the original signal. Different from most of the speaker recognition systems which use delta features to capture the dynamic properties of speech signals, here only static MFCC features are used, but frames within a window are concatenated to form a 'super frame' to encode the dynamic information. The window size is set to 9 frames (4 frames before and after the current frame, respectively). Note that the clipped speech and the original speech are in the same length, hence no alignment between the two data streams are required.

The DNN structure involves 3 hidden layers with each layer consisting of $1,200$ units. The training objective function is the mean square error between the MFCC features of the reconstructed speech and the original one, and the stochastic gradient descend (SGD) algorithm is applied to conduct the training. The learning rate starts from a relatively large value (0.008), and then is gradually shrunk by halving the value whenever no square error reduction on the development set is obtained. The training stops when the square error reduction on the development set is getting small (0.001 as the threshold). Neither momentum nor regularization has been used, and no pre-training is employed since we do not observe clear advantage by involving these techniques. Once the DNN has been trained, the MFCC features of clipped speech are fed into the DNN and the MFCC features of the reconstructed speech are read out from the DNN outputs.

5.3 Experimental results

We use the same data set and configurations as in the previous sections, with the exception that the clipped data are reconstructed by the DNN model, instead of simply discarded or are left as they are. In this experiment, the DNN model is trained with the development set (the Fisher corpus) that has been used to train the UBM and the loading matrix of the i-vector system. Again, the clipped speech data are generated by setting the clipping rate at different levels.

Two DNN models are investigated: (1) the clipping rate in DNN training is the same as that in the clipping reconstruction when testing; (2) the DNN is trained by collecting clipped speech at various clipping rates, resulting in a universal DNN-based clipping reconstructor. These two models are denoted by 'CR-dependent' and 'CR-independent' respectively ('CR' stands for 'clipping rate').

First we test the scenario where only the test utterances are clipped and thus reconstruction is required. The results are shown in Fig. 20. We observe that with the DNN-based reconstruction, the performance is considerably improved when the clipping is aggressive. If the clipping is not such aggressive, the reconstruction does not help, but importantly, it does not degrade the performance anyway. Interestingly, the CR-independent DNN shows similar performance as the CR-dependent

DNN. This is an interesting result and it suggests that a CR-independent model may be sufficient for handling various clipping rates.

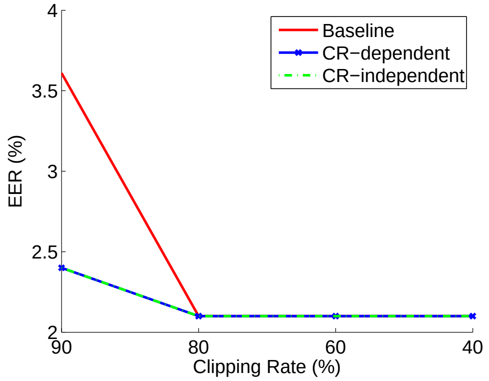

Figure 20 ERR results with DNN-based clipping reconstruction. Only the test utterances are clipped and reconstructed.

In the second scenario, both the enrollment utterances and test utterances are clipped. We apply the DNN-based reconstruction either to both of the enrollment and test utterances, or just to the test utterances. The results are shown in Fig. 21 and Fig. 22, respectively, for these two settings. We observe similar patterns as in Fig. 20 that the reconstruction helps most too heavily clipped speech. The CR-dependent reconstruction seems always stable and the performance does not degrade with weakly-clipped speech. The CR-independent reconstruction may lead to marginal performance degradation with weakly-clipped speech, but in general it works well.

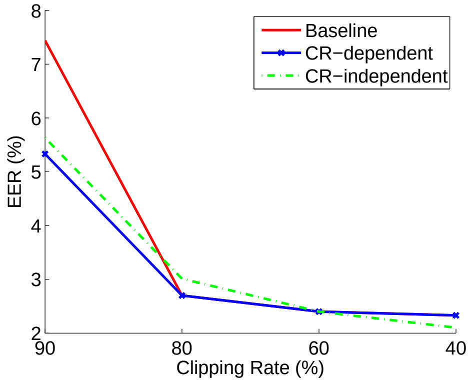

Figure 21 EER results with DNN-based clipping reconstruction. Both the enrollment and test utterances are clipped and reconstructed.

6 Conclusions

This paper studies the phenomenon of speech clipping, a common signal corruption in practical speaker recognition. We investigate the impact of speech clipping on

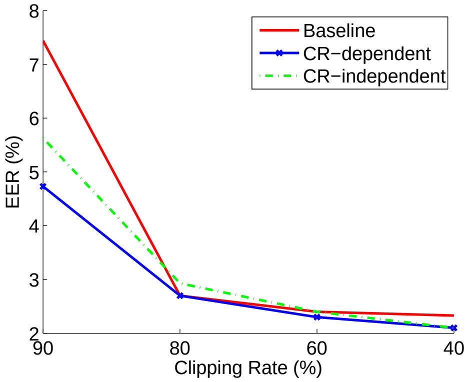

Figure 22 EER results with DNN-based clipping reconstruction. Both the enrollment and test utterances are clipped but only the test utterances are reconstructed.

two speaker recognition systems that are based on the GMM-UBM framework and the i-vector model, respectively, and find that the i-vector system performs much more robustly against clipping than its GMM-UBM counterpart. In addition, we propose a simple yet effective clipping detection approach based on the distribution of sample values, and utilize the detection result to remove clipped speech segments. The experimental results show that clipped speech still contains speaker information and it is better to retain them in practical systems. Finally, a clipped speech reconstruction approach is proposed based on the DNN model. The results show that the proposed approach can lead to considerable performance improvement for speaker recognition, particularly for heavily-clipped speech data. Future work will study complex distortions, such as clipping coupled with noises, and study better DNN structures, particularly the recurrent DNN, which is supposed to be more suitable to learn the dynamic properties of clipped speech, leading to a more effective clipped speech reconstruction.

Acknowledgements

This work was supported by National Basic Research Program (973 Program) of China under Grand No. 2013CB329302 and the National Science Foundation of China (NSFC) under the project No. 61371136 and No. 61271389.

References

1. William M Campbell, Joseph P Campbell, Douglas A Reynolds, Elliot Singer, and Pedro A Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech & Language*, vol. 20, no. 2, pp. 210–229, 2006.
2. Frédéric Bimbot, Jean-François Bonastre, Corinne Fredouille, Guillaume Gravier, Ivan Magrin-Chagnolleau, Sylvain Meignier, Teva Merlin, Javier Ortega-García, Dijana Petrovska-Delacrétaz, and Douglas A Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 2004, pp. 430–451, 2004.
3. James M Kates and Linda Kozma-Spytek, "Quality ratings for frequency-shaped peak-clipped speech," *The Journal of the Acoustical Society of America*, vol. 95, no. 6, pp. 3586–3594, 1994.
4. JOSEPH CR Licklider, "Effects of amplitude distortion upon the intelligibility of speech," *The Journal of the Acoustical Society of America*, vol. 18, no. 1, pp. 249–249, 1946.
5. Thomas R Crain and Dianne J Van Tasell, "Effect of peak clipping on speech recognition threshold," *Ear and hearing*, vol. 15, no. 6, pp. 443–453, 1994.
6. Srdjan Kitic, Laurent Jacques, Nilesh Madhu, Michael Peter Hopwood, Ann Spriet, and Christophe De Vleeschouwer, "Consistent iterative hard thresholding for signal declipping," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013*. IEEE, 2013, pp. 5939–5943.
7. Mark J Harvilla and Richard M Stern, "Least squares signal declipping for robust speech recognition," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

8. Lawrence Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
9. Jeremy Bradbury, "Linear predictive coding," *Mc G. Hill*, 2000.
10. AJEM Janssen, R Veldhuis, and L Vries, "Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 34, no. 2, pp. 317–330, 1986.
11. Ivan Selesnick, "Least squares with examples in signal processing," .
12. Amir Adler, Valentin Emiya, Maria G Jafari, Michael Elad, Rémi Gribonval, and Mark D Plumbley, "Audio inpainting," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 922–932, 2012.
13. Abdelhakim Dahimene, Mohamed Noureddine, and Aarab Azrar, "A simple algorithm for the restoration of clipped speech signal.," *Informatica (Slovenia)*, vol. 32, no. 2, pp. 183–188, 2008.
14. Mathieu Lagrange, Sylvain Marchand, and Jean-Bernard Rault, "Long interpolation of audio signals using linear prediction in sinusoidal modeling," *Journal of the Audio Engineering Society*, vol. 53, no. 10, pp. 891–905, 2005.
15. Martin Dietz, Lars Liljeryd, Kristofer Kjorling, and Oliver Kunz, "Spectral band replication, a novel approach in audio coding," in *Audio Engineering Society Convention 112*. Audio Engineering Society, 2002.
16. Paris Smaragdis, Bhiksha Raj, and Madhusudana Shashanka, "Missing data imputation for spectral audio signals," in *IEEE International Workshop on Machine Learning for Signal Processing, 2009. MLSP 2009*. IEEE, 2009, pp. 1–6.
17. Manuel Moussallam, Pierre Leveau, and SM Aziz Sbai, "Sound enhancement using sparse approximation with speclets," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), 2010*. IEEE, 2010, pp. 221–224.
18. Colin Perkins, Orion Hodson, and Vicky Hardman, "A survey of packet loss recovery techniques for streaming audio," *Network, IEEE*, vol. 12, no. 5, pp. 40–48, 1998.
19. Hadas Ofir, David Malah, and Israel Cohen, "Audio packet loss concealment in a combined mdct-mdst domain," *Signal Processing Letters, IEEE*, vol. 14, no. 12, pp. 1032–1035, 2007.
20. Simon JD Prince and James H Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *ICCV'07*. IEEE, 2007, pp. 1–8.
21. Pavel Matvejka, Ondvrej Glembek, Fabio Castaldo, Md Jahangir Alam, Oldvrich Plchot, Patrick Kenny, Lukávs Burget, and Jan vCernocky, "Full-covariance ubm and heavy-tailed plda in i-vector speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011*. IEEE, 2011, pp. 4828–4831.
22. Laurens Van Der Maaten, "Accelerating t-sne using tree-based algorithms," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3221–3245, 2014.
23. Laurens van der Maaten and Geoffrey Hinton, "Visualizing non-metric similarities in multiple maps," *Machine learning*, vol. 87, no. 1, pp. 33–55, 2012.
24. Laurens Maaten, "Learning a parametric embedding by preserving local structure," in *International Conference on Artificial Intelligence and Statistics*, 2009, pp. 384–391.
25. Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 2579-2605, pp. 85, 2008.
26. Li Deng and Dong Yu, *DEEP LEARNING: Methods and Applications*, NOW Publishers, January 2014.
27. Hervé Bourlard and Nelson Morgan, "Hybrid HMM/ANN systems for speech recognition: Overview and new research directions," in *Adaptive Processing of Sequences and Data Structures*, pp. 389–417. Springer, 1998.
28. Hynek Hermansky, Daniel PW Ellis, and Sangita Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2000*, 2000, pp. 1635–1638.
29. George E Dahl, Dong Yu, Li Deng, and Alex Acero, "Large vocabulary continuous speech recognition with context-dependent DBN-HMMs," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011*, 2011, pp. 4688–4691.
30. Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.
31. Z. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted boltzmann machines for statistical parametric speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013*, 2013, pp. 7825–7829.
32. Z. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted boltzmann machines and deep belief networks for statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2129–2139, 2013.
33. P. Hamel and D. Eck, "Learning features from music audio with deep belief networks," in *Proceedings of International Symposium on Music Information Retrieval (ISMIR)*, 2010.
34. E. Batternberg and D. Wessel, "Analyzing drum patterns using conditional deep belief networks," in *Proceedings of International Symposium on Music Information Retrieval (ISMIR)*, 2012.
35. Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. of the 25th international conference on Machine learning*, 2008, pp. 1096–1103.
36. Andrew L Maas, Quoc V Le, Tyler M O'Neil, Oriol Vinyals, Patrick Nguyen, and Andrew Y Ng, "Recurrent neural networks for noise reduction in robust ASR," in *Proc. of Interspeech*, 2012, pp. 22–25.
37. Xiao-Lei Zhang and Ji Wu, "Deep belief networks based voice activity detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697–710, April 2013.
38. A. van den Oord, S. Dieleman, and B. Schrauwen, "Deep content-based music recommendation," in *Proceedings of Neural Information Processing Systems (NIPS)*, 2013.

39. P Kenny, V Gupta, T Stafylakis, P Ouellet, and J Alam, "Deep neural networks for extracting baum-welch statistics for speaker recognition," .
40. Yun Lei, Nicolas Scheffer, Luciana Ferrer, and Mitchell McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014*, 2014.