Nonlinear waveform distortion: Assessment and detection of clipping on speech data and systems

John H.L. Hansen*, Allen Stauffer, Wei Xia
Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering, University of Texas at Dallas, Richardson, TX, USA

Abstract

Speech, speaker, and language systems have traditionally relied on carefully collected speech material for training acoustic models. There is an enormous amount of freely accessible audio content. A major challenge, however, is that such data is not professionally recorded, and therefore may contain a wide diversity of background noise, nonlinear distortions, or other unknown environmental or technology-based contamination or mismatch. There is a crucial need for automatic analysis to screen such unknown datasets before acoustic model development training, or to perform input audio purity screening prior to classification. In this study, we propose a waveform based clipping detection algorithm for naturalistic audio streams and examine the impact of clipping at different severities on speech quality measurements and automatic speaker recognition systems. We use the TIMIT and NIST SRE08 corpora as case studies. The results show, as expected, that clipping introduces a nonlinear distortion into clean speech data, which reduces speech quality and performance for speaker recognition. We also investigate what degree of clipping can be present to sustain effective speech system performance. The proposed detection system, which will be released, could contribute to massive new audio collections for speech and language technology development (e.g. Google Audioset (Gemmeke et al., 2017), CRSS-UTDallas Apollo Fearless-Steps (Yu et al., 2014) (19,000 h naturalistic audio from NASA Apollo missions)).

Keywords
Audio clipping; Speech quality assessment; Non-linear distortion; Speaker recognition

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

*Corresponding author. John.Hansen@utdallas.edu (J.H.L. Hansen).

CRediT authorship contribution statement

John H.L. Hansen: Supervision, Conceptualization, Resources, Writing – review & editing, Project administration, Funding acquisition. Allen Stauffer: Conceptualization, Methodology, Software – initial version, Validation, Investigation. Wei Xia: Conceptualization, Methodology, Software, Analysis, Investigation, Writing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
1. Introduction

The formulation of advanced speech and language technology in the past has historically relied on carefully organized data collection, usually under regulated laboratory conditions. However, nowadays a vast amount of audio data appears daily on the web, data warehouses for call centers, and personal devices such as smart phones. Due to the cost of collecting organized and focused speech/language corpora, researchers have turned to naturalistic audio material in order to reduce both cost and collection time, as well as increase the diversity of speakers, languages, and vocabulary or topic content. While there is a plethora of audio material available, the non-uniformity and recording impurities could raise great concerns regarding the viability of resulting algorithms.

Audio peak clipping occurs when the volume of the captured audio signal exceeds the input range of the microphone’s pre-amplifier, or the audio data is recorded without an appropriate input automatic gain control (AGC). Portions of the signal above the maximum voltage would be clamped to the maximum value of the signal when it passes through the analog to digital (A/D) converter. The loss of high amplitude samples introduces a non-linear distortion in the form of odd harmonics in high frequencies, resulting in audible artifacts in the recorded audio.

In order to detect such distortions, Aleinik and Matveev (2014) presented a histogram method to estimate the level of signal clipping. Ding et al. (2006) investigated the effects of temporal clipping on perceived speech quality. They proposed a non-intrusive algorithm based on the clipping statistics to predict speech quality. Temporal clipping can also occur as a result of voice activity detection (VAD) or echo cancellation where comfort noise is used in place of clipped speech segments. Authors in Eaton and Naylor (2013, 2014), Bie et al. (2015) and Deng et al. (2013) estimated the clipped samples and the original signal level with histogram analysis. Tachioka et al. (2014) analyzed the relationship between clipping level and automatic speech recognition performance and showed the explicit relationship between SNR and clipping level. Hines et al. (2013, 2016) presented a novel non-reference measure that uses a modular design that helps pinpoint the reason for degradations in addition to quantifying their impact on speech quality. Maymon et al. (2013) proposed two iterative approaches based on the band-limited assumption and autoregressive model to recover clipped speech signals. Harvilla and Stern (2014, 2015) introduced a de-clipping algorithm based on constrained least-squares minimization, where the constrained blind amplitude reconstruction algorithm interpolates missing data points such that the resulting function is smooth while ensuring the inferred data falls into a legitimate range.

For clipping, it is possible to suggest two alternative definitions for impacting clip distortion on clean speech. This includes: (i) Clip-by-Sample over time: in this definition, a clipping rate of 10% implies a clipping threshold is set, such that 10% of the overall time-domain samples are clipped; (ii) Clip-by-Amplitude: in this definition, the clipping threshold is set simply as the portion of the maximum waveform amplitude to be hard limited. In this study, we use “Clip-by-Sample over time” as our definition, arguing that this is more appropriate since Clip-by-Amplitude will not reflect the actual number of samples that are distorted,
especially if there are a few high amplitude samples such as impulsive sounds over an entire audio stream.

To illustrate the differences in clipping distortion on an audio stream, we show the difference and a mapping comparison of the two definitions - “Clip-by-Sample over time and Clip-by-Amplitude” in Figs. 1 and 2. We observe that the numerical range of the Clip-by-Amplitude definition is larger than that if we use the Clip-by-Sample over time definition. If there is a large amplitude random noise, though the clipping rate is high if we use the Clip-by-Amplitude definition, a large amount of the time samples are not actually affected. This is because of the non-stationary nature of the speech signal, especially for nasals and fricatives, which contain many low amplitude parts. If we use the Clip-by-Amplitude definition in this case, only when the clipping rate is very large, will the clipping influence the speech signal across different time frames. Due to these reasons, we decide to define the clipping rate as the percentage of clipped samples over time.

Given the ever-increasing availability of speech and audio data in the field, the speech community needs improved speech tools to better characterize and understand issues in “found” data. In this study, we restrict the potential acoustic issue to waveform peak clipping, which occurs when audio data is recorded at an improperly adjusted gain setting and without an input AGC, or when the AGC cannot respond quickly enough to suppress impulsive audio events.

The focus of this study is on assessing and detecting audio waveform peak clipping, as well as the overall impact on automatic speech-based systems. In addition to defining types of clipping, we also discuss possible causes. In later sections, we demonstrate the impact of clipping on automatic speaker identification systems, provide audio quality measurement analysis, and report results of a study of the presence of clipping in datasets used by the speech community. We will also discuss the specifics of the proposed ClipDaT algorithm used for this analysis, which will be distributed for general usage to the community. This study, however, will not address clipping repair of speech which should be addressed in subsequent studies.

1.1. Clipping overview

Audio peak clipping occurs when the volume of the audio signal being recorded exceeds the input voltage range of the microphone’s pre-amplifier given the current gain for analog-to-digital (A/D) conversion. When this occurs, the pre-amplifier voltage becomes saturated and unable to provide an accurate discrete representation for reliable A/D conversion. This causes the “peak” of the signal to not be reproduced by the pre-amplifier, which means this “peak” as well as all portions of the signal above the maximum voltage of the pre-amplifier will be clamped to the maximum of the signal as it passes through the A/D converter. Therefore the natural shape of the speech waveform is not properly represented in the discrete signal; instead, a plateau appears, and the information contained in the higher amplitude samples is lost. The manifestation of this loss of data and the introduction of this plateau shape comes in the form of non-linear distortion, especially odd harmonic distortion in higher frequencies, resulting in audible artifacts in the recorded audio. Thomas and Niederjohn (1970) initially exploited this phenomenon using their infinite-peak-clipping...
method to increase the perceived intelligibility of speech prior to transmission across noisy channels, where low energy consonant sections are greatly amplified, improving their corresponding signal-to-noise ratio.

Clipping may be encountered or introduced into audio in a number of situations. One common form is when the microphone pre-amplifier gain sets too high for the input speech. In most studio recording situations, the gain is fixed for the entire recording, thus the clipping present within a recording session will be consistent throughout the recording. Clipping in this situation is relatively easy to detect since the gain remains constant. For large studio recordings, however, this situation can become a common occurrence if the gain is not adjusted for each participant, or set once at the start of each recording session but not monitored throughout, or manually verified for quality post-recording.

In contrast to a fixed-gain studio-recording situation, most telephony systems and recording devices incorporate an automatic gain control (AGC) system. The AGC automatically adjusts the input gain of a recording device in response to the changing volume of the input signal; attenuating when larger amplitude sounds are encountered and boosting during quieter periods so as to maximize the available dynamic range of the A/D converter. The addition of an AGC system allows for low quantization noise by maintaining a high gain when possible, but reducing it when the volume rises, theoretically before issues such as clipping occur. These systems greatly mitigate many clipping-inducing situations, however, due to limitations in how fast systems can respond to amplitude changes or a slow attack time (the time a system respond to change in input amplitude), and physical limitations in how low the gain can be adjusted, conditions are ripe for a range of clipping scenarios to occur.

Naturalistic or uncontrolled recordings can contain clipping caused by many different sources, including loud but short impulse-type noises, such as hard impact sounds, construction tool noise sources, gunshots, and other unseen events such as screaming, sirens, fire, and car horns. These examples are more likely to be found in spontaneous, real-world data collections. A subject coughing, sneezing, touching a microphone, or adjusting their clothing, if a lapel-mounted microphone is being used, can all cause clipping to be introduced into the recording. Speech systems will obviously be most affected by clipping that occurs within the waveform sections of actual speech. With this in mind, the results presented in the next section that refers to the impact of clipping on such systems are performed with data known to be clean of these unseen events, to eliminate the variability this type of data may have on results. We present the details of the data in Section 1.1. Additional sources of clipping can also include artifacts introduced by passing audio through transmission channels with limited amplitude bandwidth. Compressing audio, converting it to other audio formats, and down-sampling can also introduce clipping after an initial recording and A/D processing. Compressed low bit rate audio formats are becoming increasingly common, especially with the already large and ever-growing popularity of streaming video websites such as YouTube, Voice over IP services such as Skype, and the inherent constraints of streaming audio and video contribute to increased opportunities for clipping to be introduced. As a first pass, we could detect the presence of clipping by determining when the audio signal reaches the +/- maximum digital representation provided
by the A/D converter. However, due to the signal re-transmission, signal data reformatting (i.e., converting from linear PCM format to μ-law/α-law or other audio compression standards), these peak clipped values could be morphed to other amplitude levels. A simple search of +/- max values would therefore miss these clipping events.

In Fig. 3, we provide a graphical illustration of the impact of clipping for three different phoneme classes: vowel /ei/, nasal /n/, and fricative /sh/, respectively. We create these visuals by isolating the three phonemes in question, and increasing the gain of each until 10% of the audio time domain samples migrate to an extreme.

Due to the higher energy found in voiced speech segments such as vowels, liquids, glides, diphthongs, there is a higher likelihood that such segments would experience clipping, while nasals and fricatives are less likely. Stop consonants could also experience clipping during the air burst portion, especially if the microphone is close to the mouth of the speaker.

The first row of Fig. 3 presents time waveforms of the /ei/, /n/, and /sh/ phonemes from left to right respectively, with blue representing the original unclipped signal, and red the corresponding clipped version. These waveforms serve as a reference to show how waveform clipping results in changes viewed in the spectrograms, and mel frequency cepstral coefficient (MFCC) plots. The second row contains spectrogram representations of the original unclipped waveforms in the second row, versus clipped waveform spectrograms in the third row. Comparing the corresponding spectrogram plots of each phoneme, we can observe that the largest impact of clipping is noise harmonics appearing in the higher frequencies of the spectrogram. This manifests as more energy within the upper regions of the spectrogram. We note that the lower frequencies, the overall shape of the speech energy information still remains, and the formant structure appears to be present and intact. This would imply that the audio is still intelligible, and that, for this level of clipping, the content is not degraded to the point that it is not still easily understood. Anecdotally, informal listening assessment supports this observation, with the speaker’s intentions very clear but with noticeable high frequency artifacts present such as pops and hisses.

Finally, the fourth row of Fig. 3 displays the overall average MFCC plots for these same three pairs of waveforms. These plots represent a small slice of the data that would be processed and used by classifiers in automatic speech-processing systems to represent speakers or phonemes. Comparisons across these plots show how different clipped versus unclipped files would be interpreted by automatic speech systems.

The overall shape of the clipped-audio MFCC vectors remains relatively similar to their unclipped counterparts. However, there is sufficient variation in individual coefficients that impact the performance of speech systems. In Section 2, we will quantify the impact of clipping on speech quality assessment and speaker recognition systems.

2. Analysis: Clipping impact on speech

In this study, clipping is artificially introduced into the original clean corpora in order to analyze the effect peak clipping has on two classes of automatic speech systems under
controlled conditions: quality measures and speaker identification. This section will describe the dataset and systems used for these experiments, and discuss the results obtained.

2.1. Generation of clipped speech data

The clipped speech data used in this phase of the study is created using data from the TIMIT Acoustic-Phonetic Continuous Speech Corpus (Garofolo et al., 0000) and NIST 2008 speaker recognition evaluation (SRE) dataset (Alvin F. Martin, 2008). The TIMIT corpus consists of a large number of speakers recorded in a studio setting while reciting ten phonetically-balanced sentences. The NIST 2008 SRE dataset is a standard dataset for experiments in text-independent speaker recognition.

In order to clip the data in a consistent and controlled manner, each file is first evaluated by the Combo-SAD speech activity detection (SAD) algorithm developed by Rix et al. (2001). This allows us to focus on the clipping that appears during recorded segments where the speaker is actually speaking. Next, we independently increase the gain for each file until reaching a pre-set percentage of time-domain speech samples, as declared by the labels output from the Combo-SAD tool in the previous step, “hit the rails” or land at the extreme +/- maximum value. Four clipped versions of TIMIT are created for these experiments, one each at 0.5%, 1%, 5%, and 10%, where these numbers represent the percentage of time-domain samples of the speech waveform that are clipped.

We artificially introduce clipping to TIMIT and SRE2008 data. This allows us to control the amount of clipping, while being certain that external factors, such as microphone type or placement, or channel conditions are not introducing additional variability into our results.

2.2. Impact of clipping

This section presents two studies on the impact of peak clipping. The first deals with how common speech quality measures react to the introduction of clipping to recorded speech data. The second addresses how automated speaker identification is impacted by clipping data.

2.2.1. Speech quality assessment—We use four speech quality measures to evaluate the clipped speech data. The first, Perceptual Evaluation of Speech Quality (PESQ) (Rix et al., 2001) is a full reference measure derived from the ITU-T P.862 recommendation regarding speech quality for telephony applications. The PESQ speech quality test takes two versions of the same signal; one clean, and one modified/degraded, and compares them. The output report from the PESQ algorithm is a mean opinion score (MOS), that ranges from -0.5 to positive 4.5, with 4.5 being the highest, representing excellent quality.

The speech-to-noise-ratio (STNR) algorithm supported by the National Institute of Standards and Technology (NIST) is based on the idea that speech segments will possess a higher energy content than frames or segments that contain only silence/noise. Expanding on this idea, by analyzing the energy distribution of frames of the audio file, the algorithm can then use these energy values to create a histogram that represents the overall distribution of the signal energy. This histogram, coupled with the assumption that voiced speech frames will contain similar energy and therefore cluster together, and likewise silence or noise-only
frames will also cluster together, allows the algorithm to estimate the distance between the speech and noise content of the bimodal histogram distribution of the audio file. The NIST STNR measure is a non-reference quality measure where higher numbers represent a higher quality measurement. This measure assumes that the speech signal energy is greater than the silence/noise frame content, and therefore assumes a positive STNR response.

Waveform Amplitude Distribution Analysis signal-to-noise ratio (WADA-SNR) is the third quality measure evaluated. Kim and Stern (2008) proposed WADA-SNR as a non-reference quality measure. Unlike NIST STNR, this approach is not based on analysis of energy, but on modeling the amplitude distribution of a speech waveform with the Gamma distribution function, while assuming that the background noise present will fit a Gaussian distribution. For WADA-SNR, higher output score values also represent higher quality for the audio recording.

Finally, analysis results are also be presented using the Blind Source Separation Evaluation (BSS Eval) toolbox (Vincent et al., 2006) to measure the Sources to Artifacts Ratio (SAR) of each clipped file compared to its corresponding clean original. This measure decomposes the signal to be evaluated into estimates of the signal’s true source, spatial distortion, interference, and artifacts, and compares each of these estimates respectively to the actual true source. In this case, the true source is the clean, unclipped audio recording, where the artifact estimate is the term we are interested in assessing. A larger output number for SAR corresponds to fewer and/or less severe artifact introductions and therefore represents a higher overall signal quality.

Table 1 presents results obtained when evaluating the data from all 630 speakers of the artificially clipped TIMIT dataset with the aforementioned speech quality measures. All four speech quality measures are clearly sensitive to the presence of clipping, with scores diminishing as the level of introduced clipping increases for each algorithm. We can see that the WADA SNR reported for the unclipped audio recordings is lower than both 0.5% and 1.0% clipped recordings, suggesting that the original TIMIT dataset may have some slight distortion. For the remainder of the results, we can observe a clear degrading trend compared to the original unclipped audio. Correlation between clipping and these quality measures are high and consistent, with the jump from 1% clipping to 5% clipping having a much larger impact on speech quality scores versus the move between 0.5% and 1% clipping.

2.2.2. Impact on human listeners—In many cases, human listening is used as the primary screening test for the presence of distortion such as clipping. In some cases, distortion may not be as perceptually noticeable even if it impacts the speech system performance. In order to measure how closely the speech quality trend demonstrated in Section 2.2.1 applies to human perception of the clipping phenomenon, we conduct a human listener test. Participants of this study are recruited from the general University of Texas at Dallas population. They range in age from 20–35 years, and include a mixture of native American English (AE) speakers as well as non-native AE speakers. All have no history of hearing loss. We present all listeners with twenty audio files with varying levels of clipping, and ask them to rate each on a mean opinion score (MOS) scale. Listeners evaluate each audio using a MOS scale with a numerical range from 1 to 5, where each value is defined
as follows: (1) Bad: very annoying, (2) Poor: annoying, (3) Fair: slightly annoying, (4) Good: Clipping/distortion is perceptible, but not annoying, (5) Excellent: Clipping/distortion is imperceptible. We present the same 20 files in the same order to each listener. The distribution of clipping throughout the 20 files is as follows; there are 4 each of 5 different levels of clipping: clean/unclipped, 0.5%, 1.0%, 5.0%, 10.0%, and 15.0%. We show the results in Table 2, which reflect the average opinions across the total 15 participants.

From the results in Table 2, it is clear that human listeners experience the same downward trend in speech quality suggested by the automatic quality measures discussed in the previous section. Human listeners are more generous with speech quality ratings than PESQ, though it is important to note that they are given whole-number score denominations to choose from, unlike PESQ which reports scores with a much higher degree of granularity, and also that each data point in the PESQ column of Table 2 represents the average of 6300 samples, while each entry in the Human-Listener MOS column consists of just 60 averaged scores. The high correlation between the results of this human listening study with scores for automatic speech quality measures reinforces the conclusions drawn in Section 2.2.1. We also note that listeners could perceive the quality gap between 1% clipped and 5% clipped audio less clearly than the difference between the 10% and 15% clipping levels.

Speaker identification (SID) refers to the problem of recognizing a speaker from an unknown speech utterance given a pool of known target speakers. The system is usually trained on these speakers beforehand, by extracting features from the audio, and using a classifier to compare a model of these features to features extracted from the unknown sample. We firstly apply a Gaussian Mixture Model (GMM) classifier with the Universal Background Model (UBM) (Reynolds et al., 2000). We perform experiments on TIMIT data to analyze the influence of clipping distortion on training, test data respectively and whether removing clipped speech regions will improve the system performance. Next, we use the state-of-the-art i-Vector and x-Vector approaches (Kenny et al., 2007; Matějka et al., 2011; Snyder et al., 2018) with a Gaussian PLDA backend (Kenny, 2010) on the NIST SRE2008 data to further investigate the clipping effect.

**GMM–UBM:** On TIMIT data, we extract MFCC features and use a GMM with UBM. Here, 12-dimensional MFCC vectors are extracted and used to calculate delta and delta–delta/acceleration vectors, which are also provided to the classifier. Cepstral mean and variance normalization is applied to the features to reduce the impact of any microphone or channel effects on the data. The GMM–UBM classifier uses 256 mixtures for the models, and the UBM is created from 250 independent TIMIT speakers, with the same eight sentences from each speaker. The remaining 380 TIMIT speakers who are disjoint from the set of speakers used in the UBM are chosen as target speakers, using the eight sentences selected for the UBM as the training data, and the remaining two open sentences as test utterances. The training data and UBM for each test are left the same, and we use the clean, unclipped data for training. Five experimental trials are performed, with test utterances ranging from clean to 10% clipped speech samples. This approach is chosen with the assumption that once a SID system has been trained, the models are unlikely to be retrained unless a new speaker or group of speakers is being added to the set of target
speakers. This setup allows us to determine the impact of varying levels of clipping on a system trained without knowledge of the presence of clipping.

We explore the impact of clipping on the test data when using clean data for training; and second, examine a case where various amounts of clipped speech data is introduced into the training phase with clean data as the test. From Fig. 4 and Table 3, we can observe there is a highly negative correlation between the amount of peak clipping in test speech and overall SID performance. Each increase in clipping results in a corresponding increase in equal error-rate, marking a decrease in the overall SID accuracy. This trend is quite significant and consistent, as with a mild 0.5% level of clipping, an overall drop in SID performance occurs, as seen in Fig. 4. This result is probably due to the nonlinear nature of clipping, and the fact that high-energy voiced segments such as vowels and diphthongs, which contain the most useful speaker-dependent information, are the first speech classes degraded if the input gain is set too high for a given recording.

Next, additional experimentation is performed to determine the detrimental impact of using clipping-contaminated audio data to train speaker models for SID. The same TIMIT audio data is used for these trials. The results presented in Fig. 5 are obtained using clean test data, and a combination blend of clean and clipped data for training speaker models. The proportion of clipped-to-clean data is indicated on the DET curve legend. It is important to note that for these plots, when the data is listed as containing a ‘clipped’ sentence, the audio recording has been re-sampled at an increased gain until 10% of the speech samples are clipped. All clipped data for these results is clipped at this severity (i.e., for the DET curve labeled “6 Clean, 2 Clipped”, each speaker’s model is trained with 6 clean, unclipped sentences and 2 sentences that have been clipped at 10%).

Fig. 5 and Table 4 present the DET curve and EER of SID results for four experimental trials showing the impact of clipping contamination on training data at different severity levels. “8 Clean” represents the same setup in the previous experiment, with subsequent results achieved by replacing 2 of the unclipped sentences with the clipped ones. This figure demonstrates a clear downward trend in EER as more training data is replaced with clipping-degraded speech. The performance impact from only replacing a quarter of the sentences with clipped audio is not severe, with EER increasing from 3.20% to 3.95%. However, once the majority of the model is derived from clipped data, SID performance decreases significantly with EER increases to 7.63%. This leads to a conclusion that while a small amount of contamination from clipping is not catastrophic, it should still be carefully checked.

We perform a third experiment to investigate the trade-off that may exist in certain practical data-scarce SID training conditions. We would like to explore which factor might be more important for SID: (i) omit precious, but known to be clipped data from the training phase, or (ii) include contaminated data with the clipping distortion for more training samples. We set up a data-scarce condition by allowing only two pristine sentences and six clipped sentences as the total amount of training audio files per speaker.
DET curves in Fig. 6 and EER results in Table 5 show the SID performance in a data-scarce training condition with only two clean sentences (4.5–6 s) per speaker. We can observe that it might be slightly helpful to include a small amount of clipped sentences in order to have more training data to fit the model in a data-scarce condition. However, as more contaminated clipped sentences are introduced for model training, the performance drops from 6.91% EER with four additional clipped sentences, to 7.63% EER when all six clipped sentences are used. It indicates the necessity of monitoring the presence of clipping. In a data-scarce condition, a small amount of contaminated data might be utilized if we need more training samples for better model generalization, but the performance will be adversely affected if the percentage of clipped data becomes high.

The final SID-related experiment is performed to determine if removing clipped regions from the input audio sequences will help or hurt the SID performance. We use the proposed ClipDaT algorithm to tag clipped locations and remove the clipped parts from the original audio before feature extraction. This approach of directly removing clipped parts is adopted as a pre-processing step so we do not have to discard an entire audio if it contains any clipping. The delta and acceleration feature components may be affected a little bit by altering the audio's natural temporal information.

Fig. 7 and Table 6 show the DET plot and EER results respectively of the SID trials seen in Fig. 5 with and without clipping region removal processing. It shows that if a large amount of data contains clipping, removing clipped parts in the audio waveform is beneficial to the system performance. However, if there is only a limited amount of clipped data, as we can see in the “6 Clean, 2 Clipped” case, the EER is almost the same after clipping removal. The overall DET curves with clipping removal reveal a better performance in many regions. It is also important to note that even with clipping region removal, the performance is still not improved close to the performance of the system on the naturally unclipped, clean data. This suggests that careful data analysis and pruning is preferable if possible, over attempting to remove such contamination after the fact.

**I-Vector with PLDA:** We further conduct experiments on the SRE08 dataset to explore the clipping influence on an i-Vector based speaker recognition system. SRE08 dataset is a standard dataset for performing experiments on speaker recognition collected by NIST. SRE08 data includes multiple microphones and recording environments, like conversational telephone speech and conversational speech data recorded with a microphone in an interview scenario. Some of the speakers in the telephone conversational data are bilingual and their evaluation data may include speech in a language other than English as well as speech in English. The microphone recorded interview data is all in English.

There are eight test conditions in total for the SRE08 data. We focus on the core short2–short3 test condition. The short2 and short3 conditions include two-channel telephone conversational excerpts of about five minutes total duration. We perform gender-dependent speaker recognition on all male data. All training and test data involve only English language telephone speech spoken by native U.S. English speakers. The speech data are sampled at 8 kHz and coded in PCM format.
We first extract 60 dimensional MFCCs, including delta and delta–delta acceleration vectors. We train a 2048-mixture Universal Background Model. We extract 400 dimensional i-Vectors and use probabilistic linear discriminant analysis (PLDA) scoring to train the speaker model. A total of 39 433 test trials are used in this evaluation.

**X-Vector with PLDA:** To examine the impact of clipping on recent neural speaker embeddings, we also train a standard x-Vector model, which is based on a Time-Delay Deep Neural Network (TDNN) architecture that computes speaker embeddings from variable-length acoustic segments. It has five layers of 1-D CNN and uses a statistical pooling layer to encode the variable-length feature input to a fixed dimensional vector, as well as a softmax output layer. The embeddings are extracted after the statistics pooling layer. We use 400 dimensional x-Vectors and a PLDA backend for the verification task on SRE08 data.

Fig. 8 shows the DET curve of our i-Vector and x-Vector speaker recognition system at 5 different clipping rates based on Clip-by-Sample. With the increase of the clipping rate, we observe that SID accuracy drops correspondingly. When the clipping rate increases from 5% to 10%, we see a clear decrease in performance. We also show the EER results in Table 7, which matches our earlier observations using the GMM–UBM model to test clipping.

Due to the non-linear characteristic of speech signal clipping, the speaker-dependent information contained in voiced segments is likely to be clipped, which suggests the main reason that performance of our presented models drops. Therefore, researchers should be cautious concerning audio clipping when designing speaker recognition systems.

### 3. Clipping detection: ClipDaT algorithm

On the first inspection, it may appear that detection of clipping within a speech audio stream would be quite straightforward, assuming that the clipping occurs at the +/- maximum level of the output A/D digital representation. However, it is possible to experience clipping at other levels if speech is initially clipped, and then re-transmitted or converted to other audio file formats, where a renormalization of the overall gain is introduced after clipping. Therefore, clipping detection requires the assessment of any constant peak data that can occur in an isolated event or over a major block of data. The proposed algorithm, entitled ClipDaT — Clipping Detection and Tagging, is investigated in order to provide effective clipping detection for naturalistic audio streams.

The ClipDaT algorithm searches for strings of consecutive audio samples whose amplitudes reach a specific +/- maximum sample found in the file, or whose amplitudes come very near to this +/- maximum. Accommodations are made to allow for phenomena observed during clipping events described below.

Fig. 9 shows the flow diagram of the ClipDaT algorithm. The first step of ClipDaT is to perform an initial pass through the audio file to determine where the +/- maximum sample values are located. During a second pass, the algorithm searches for these +/- maximum values. Once one of these extreme samples is encountered, the immediately succeeding samples are analyzed in turn to decide if the extreme sample is the start of a string of clipped samples, or an isolated peak itself. If two consecutive extreme samples are found,

*Speech Commun. Author manuscript; available in PMC 2022 June 30.*
the algorithm considers this to be a clipping event. For the following samples to be deemed part of this clipping event, they must remain close to this extreme value, we set the threshold as 99.5% of the +/- maximum values found within the file. We adopt this approach after analyzing a number of actual clipped speech files and discover that a high amplitude input that causes clipping may not result in a steady, flat waveform. We can observe some slight variations in amplitude values once the preamplifier has been saturated. Fig. 10b provides an example of this phenomenon.

These unusual clipping circumstances are probably due to the manifestation of the behavior of the pre-amplifier, or the result of the A/D converter. In any event, the effects on the resulting waveform and speech systems can be significant as can be seen in this corresponding waveform. Additionally, a certain number of samples are allowed to dip below this threshold, while still being included as part of a sequence of clipped samples, so long as it is followed by a sample that reaches the maximum or minimum. Arbitrarily, three samples are allowed that do not meet the 99.5% threshold; this count is reset each time a maximum/minimum sample is encountered, allowing for an additional three below-threshold samples between extreme amplitudes to be part of a continuous clipped section. Fig. 10c displays an example of a waveform where this leniency portion of the algorithm allows it to be more sensitive to clipping events.

4. Corpora evaluation and benchmarking

Having formulated the ClipDaT algorithm in the previous section, in this section we employ the technique to assess both publicly available speech corpora and private speech datasets that are employed in the community. Researchers in the speech and language processing field generally assume that organized audio data collections available for public use in research would be entirely free of waveform clipping. In fact, clipping can and typically is present in most corpora. There is no recognized standard as to the amount of clipping which is acceptable within a corpus, nor does there exist a publicly available tool for detecting and assessing such conditions.

We present findings stemming from a study of clipping assessment in current popular corpora in the speech processing research community. The focus here is to illustrate how the ClipDaT algorithm can be used as an analysis tool that can both allow researchers to view characteristics of a corpus before they decide to use it for their research, or as a quality control and sanity check toolkit for data creators or publishers to re-affirm that their recordings are properly performed.

For each corpus, We employ the ClipDaT algorithm presented in Section 3 across all files. We then use the output clip tags for each speech/audio file to construct a summary visual representation, where the top of each figure contains an overall corpus pie chart that details the spread of the corpus with respect to unclipped audio files (green), recordings containing clipping (blue), and the actual amount of time taken up by clipped samples (red). The time cited for clipped audio (red) is included within the time listed for “Files with Clipping” (blue) for each graphic. These statistics are provided on the original audio files without voice activity detection.
Next, we use the collection of clip tags from all corpus files to build a composite visualization to highlight the extent of clipping across the corpus. For each corpus, we show the entire speech corpus in a pie chart on the top, including files that contain clipping (blue), the total amount of clipped samples (red), and clean files without clipping (green). The lower plot is a graphical representation of how clip events (red) are distributed throughout the corpus. The file with the largest amount of clipping is plotted to the left and the least amount of clipping is plotted to the right. Clip events are represented by red pixels and white represents clean unclipped audio. These images help visualize the location and severity of clipping contamination, allowing us to analyze whether the problem is isolated to just a few files or a widespread issue that requires attention.

In this study, specifics regarding the type of corpus, speaker size, speech data duration, language, and context/score are noted. We summarize overall benchmark scores with details presented in subsequent sub-sections.

4.1. Pan-Arabic Dialect corpus

The Pan-Arabic Dialect speech corpus has been used previously for research in the area of dialect identification (Lei and Hansen, 2011). It consists of a total of 500 speakers contributing to five regional dialects of Arabic, with an equal number of speakers in each. The dialectal regions represented in this dataset are the United Arab Emirates, Egypt, Iraq, Palestine, and Syria. Each recording session employs a lapel microphone for each of the two speakers engaged in conversation. A table microphone for each conversation participant is also available. Four distinct conversations are recorded within each session (see Fig. 11).

With a total of 5.64 s of clipped samples present in the 22.22-hour corpus, less than 19% of the total number of files are corrupted with some clipping. All files in the BLUE portion contain at least some clip samples, these files are ranked in order from the largest amount of clipping (left side) to the least amount of clipping (right side) in the lower plot. The cluster of dark pixels on the left side of the lower graphic shows that a few files contain the majority of the clipped speech, with the larger portion of the 19% of data files having only random isolated clip events. In essence, the white space reveals that only a small number of files contain the majority of the corpus clipping. Though the Pan-Arabic corpus is a naturalistic two-way conversation-based corpus, where speech content is spontaneous, the ClipDaT profile confirms it is a clean corpus with very small amounts of total clipping.

4.2. Noisy Telephone corpus

The Noisy Telephone (NoTel) speech corpus has also previously been used to perform experiments with respect to noise-robust SID (Godin et al., 2013). It consists of telephone speech recordings totaling 11,027 sessions in various naturally noisy environments such as a cell phone inside a moving vehicle, standing at a roadside or public place with a cell phone, or talking on a land-line phone within a noisy office environment. The telephone devices used were not regulated throughout the collection of the corpus, and both cell phone and land-line transmission channels were used.

Clipping content of the NoTel corpus consists of just 12.82 s spread throughout 2700 min (45 h) of audio, with 27% of the files in the corpus containing some clipping (see Fig. 12).
However, the ClipDaT distribution graphic shows that the bulk of this clipping is contained within a very small subset of these files, rather than spread evenly throughout the afflicted sessions, despite the noisy nature of the data. As such, while this is a noisy, naturalistic corpus, it is quite free of actual nonlinear clipping distortion.

4.3. Speech Under Simulated and Actual Stress corpus

The Speech Under Simulated and Actual Stress (SUSAS) corpus was released in 1999, consisting of over 16,000 single word speech utterances from 32 speakers recorded in both real and simulated stressful conditions (Hansen et al., 1997). These conditions have 11 simulated “stress” conditions, including fast/slow and loud/soft speech, angry emotion, stress-induced with computer response tasks, Lombard effect speech due to speech produced in noise, and speech produced during performance of two amusement park rides (i.e., Free Fall, Scream Machine roller coaster ride) (see Fig. 13).

Given the relatively small size of this speech corpus, just 3.88 h, the 29.22 s of clipped audio is noteworthy. However, it is not completely surprising when considering the conditions under which the speech was recorded, which includes screaming on a roller coaster. The clipping is fairly well distributed throughout the files which contain it, with clear band patterns throughout the distribution graphic that imply that certain recording scenarios were more prone to induce clipping than others. The ClipDaT profile suggests clipping is present in a large number of the actual stress files.

4.4. Switchboard 2 corpus

Phase 2 of the Switchboard 2 is another telephone speech corpus, made up of 4472 five-minute conversations from 679 speakers totaling 372 h (Graff et al., 0000). Participants were recruited from a variety of college campuses throughout the United States, and directed to discuss a specific topic during each call. All speech is natural/spontaneous.

Switchboard 2 Phase 2 is a very large corpus. Only 11% of the files have any form of clipping, with just 17.59 s of clipping spread throughout the clipped sections. The majority of this clipping is contained in just a small portion of the clipped files, as previously observed with the Pan-Arabic and NoTel corpora. The ClipDaT profile in Fig. 14 shows this corpus is essentially free of clipping in most files.

4.5. Speech in Noisy Environments (SPINE) corpus

We analyze both SPINE-1 and SPINE-2 with ClipDaT to produce the characteristic profile in Fig. 14. For SPINE-2, both the unprocessed (original microphone recordings) and processed (using various voice coders Hansen et al., 2001) are both included in the analyzed audio. The audio consists of conversational recordings of 64 two-person collaborative tasks. The recordings were made in controlled acoustic room environments with pre-recorded military noise types (6 in SPINE-1, 8 in SPINE-2) played through speakers at the time of speech recording.

The ClipDaT graphic profile from Fig. 15 demonstrates that the SPINE-1 and 2 corpora contain an inconsequential amount of clipped data; less than one second total. The lower
graphic above shows that even the files that are contaminated have only minor isolated clip events and no heavy concentrations of clipping; a small subset of the afflicted files contain the majority of the clipping.

4.6. Robust Automatic Transcription of Speech corpus

The DARPA Robust Automatic Transcription of Speech (RATS) corpus is comprised of a large amount of primarily telephony audio from various languages degraded by transmitting the audio through 8 different radio transmission conditions. Further details about the corpus and channel conditions used can be found in Kabir et al. (0000). The corpus contains speech in Arabic, Farsi, Dari, Urdu, and Pashtu. RATS actually encompasses a larger collection of audio released from LDC. We perform this evaluation on the E63 release (see Fig. 16).

Clipping is a serious affliction for the RATS corpus. The binary nature of the ClipDaT distribution graphic implies that the majority of the clipping is limited to certain channel conditions, and sparse otherwise, but the heavy concentration of clipping within files that have clipping present on the left side of the graphic show that it is a serious issue that must be considered when performing research experiments using this data. Table 8 summarizes six corpora evaluated with ClipDaT.

5. Discussion and conclusions

In this study, we have investigated the causes and impact of nonlinear distortion represented by clipping. An analysis of clipping effects for subjective and objective speech quality assessment was presented based on four popular speech quality measures and a human listening test. We also showed the impact of clipping on both training and test material for speech technology using a speaker recognition platform. Finally, using the proposed ClipDaT algorithm, we analyzed the clipping effect of six corpora available to the speech community.

In general, clipping has a degrading effect on the accuracy of speech systems including automatic speaker identification. The effects range from a trivial drop in accuracy when only a small number of files or short duration of audio is affected, to a significant performance loss when clipping contamination is widespread. We should be aware of this phenomenon, and be responsible for deciding whether or not the severity of clipping within the data being used is high enough to require that it be addressed. With this in mind, we suggest the following recommendations based on the results presented in this study as best practices for the speech and language community.

5.1. Recommendation #1

With files having a level of 1% or below clipping-to-speech ratio (CSR), the impact on SID accuracy from clipping will be minimal. As the amount of clipping rises to 5% CSR, the demonstrated impact, and therefore cause for concern, is much higher. At 10% CSR, the file should no longer be considered a reliable sample. These files should likely be omitted from the experiment, or in cases of data scarcity, processed to remove specific clipped sections, or an attempt is made to restore or repair clipped portions of the signal. However, it is noted that the accuracy improvement gained from removing sections of clipping is not significant
enough to fully mitigate the performance drop realized by using contaminated versus clean data.

5.2. Recommendation #2

If only a small number of files within a dataset suffer from clipping contamination, the negative performance impact will likewise be minimal. Fig. 5 and Table 2 show a small, less than 1%, drop in EER, occurs when 25% of the training data used suffers from clipping, which becomes dramatic when more than half of the audio has clipping. In a scarce data situation, a small portion of clipped files could be allowed without dire consequences, but if the total contamination rises over 25% of the files, unless the CSR is very low for all files involved, one should address this problem before further proceeding with the SID system.

5.3. Conclusions

This study has therefore explored a number of issues where clipping contamination can cause a performance impact for speech quality assessment and speaker recognition systems. It is noted that CRSS-UTDallas will release the ClipDaT toolkit to the speech and language processing community to provide automatic detection and tag clipping occurrences, allowing for both speech corpus creators to detect and resolve issues early on, as well as for researchers to be aware of the presence and decide how to proceed with the affected speech files. A survey of clipping in commonly used speech corpora was presented, revealing that while it may be believed based on listener perception that nonlinear distortion such as clipping is not a common occurrence in popular available speech corpora, it is actually present to varying degrees in almost all available speech material. Finally, speech and language technology is currently moving towards the use of more open-domain and publicly available speech material for system development, especially for training acoustic models in most speech/language classification tasks. Due to this, the community needs an effective, publicly available tool to assess this distortion. We believe that releasing the ClipDaT toolkit to the community will help contribute to both speech corpus developers as well as speech/language researchers by providing them a better understanding of their data, as well as offering more confidence in system development based on audio content without clipping distortion.

Acknowledgments

✩ This project was funded in part by AFRL under contract FA8750–15–1–0205 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen.

References

Aleinik S, Matveev Y, 2014. Detection of clipped fragments in speech signals. Int. J. Electr. Electron. Sci. Eng. 8 (2), 74–805, (0000).
Martin Alvin F., C.S.G., 2008. The nist year 2008 speaker recognition evaluation plan. http://www.nist.gov/speech/tests/spk/2008/index.htm.
Bie F, Wang D, Wang J, Zheng TF, 2015. Detection and reconstruction of clipped speech for speaker recognition. Speech Commun. 72, 218–231.

*Speech Commun. Author manuscript; available in PMC 2022 June 30.*
Deng F, Bao C.-c., Bao F, 2013. Clipping detection of audio signals based on kernel fisher discriminant. In: Signal and Information Processing (ChinaSIP), IEEE China Summit & International Conference on. pp. 99–103.

Ding L, Radwan A, El-Hennawy MS, Goubran RA, 2006. Measurement of the effects of temporal clipping on speech quality. IEEE Trans. Instrum. Meas. 55 (4), 1197–1203.

Eaton J, Naylor PA, 2013. Detection of clipping in coded speech signals. In: Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European, pp. 1–5.

Eaton J, Naylor PA, 2014. Noise-robust detection of peak-clipping in decoded speech. In: Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on. pp. 7019–7023.

Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS, 0000. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1–1.1, NASA STI/Recon technical report n 93.

Gemmeke JF, Ellis DPW, Freedman D, Jansen A, Lawrence W, Moore RC, Plakal M, Ritter M, 2017. Audio set: An ontology and human-labeled dataset for audio events. In: Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on, New Orleans, LA.

Godin KW, Sadjadi SO, Hansen JH, 2013. Impact of noise reduction and spectrum estimation on noise robust speaker identification. In: INTERSPEECH. pp. 3656–3660.

Graff D, Walker K, Canavan A, 0000. Switchboard-2 phase ii. https://catalog.ldc.upenn.edu/LDC99S79.

Hansen JH, Bou-Ghazale SE, Sarikaya R, Pellom B, 1997. Getting started with susas: a speech under simulated and actual stress database. In: Eurospeech, 97. pp. 1743–1746.

Hansen JH, Sarikaya R, Yapanel U, Pellom B, 2001. Robust speech recognition in noise: an evaluation using the spine corpus. In: Seventh European Conference on Speech Communication and Technology.

Harvilla MJ, Stern RM, 2014. Least squares signal declipping for robust speech recognition. In: INTERSPEECH. pp. 2073–2077.

Harvilla MJ, Stern RM, 2015. Efficient audio declipping using regularized least squares. In: Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on. pp. 221–225.

Hines A, Skoglund J, Kokaram AC, Harte N, 2013. Monitoring the effects of temporal clipping on voip speech quality. In: INTERSPEECH.

Hines A, Skoglund J, Kokaram AC, Harte N, 2016. Monitoring voip speech quality for chopped and clipped speech. Communications.

Kabir A, Giurgiu M, Barker J, 0000. The robust automatic transcription of speech corpus. https://catalog.ldc.upenn.edu/ldc2015s02.

Kenny P, 2010. Bayesian speaker verification with heavy-tailed priors. In: Odyssey. p. 14.

Kenny P, Boulianne G, Ouellet P, Dumouchel P, 2007. Joint factor analysis versus eigencannels in speaker recognition. IEEE Trans. Audio, Speech Language Process. 15 (4), 1435–1447.

Kim C, Stern RM, 2008. Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. In: INTERSPEECH. pp. 2598–2601.

Lei Y, Hansen JH, 2011. Dialect classification via text-independent training and testing for arabic, spanish, and chinese. IEEE Trans. Audio Speech Language Process. 19 (1), 85–96.

Matějka P, Glémek O, Castaldo F, Alam MJ, Plchot O, Kenny P, Burget L, Černocky J, 2011. Full-covariance ubm and heavy-tailed plda in i-vector speaker verification. In: Acoustics, speech and signal processing (ICASSP). In: IEEE International Conference on. pp. 4828–4831.

Maymon S, Marcheret E, Goel V, 2013. Restoration of clipped signals with application to speech recognition. In: INTERSPEECH. pp. 3294–3297.

Reynolds DA, Quatieri TF, Dunn RB, 2000. Speaker verification using adapted gaussian mixture models. Digit. Signal Process. 10 (1–3), 19–41.

Rix AW, Beerends JG, Hollier MP, Hekstra AP, 2001. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In: Acoustics, Speech, and Signal Processing (ICASSP), IEEE International Conference on, Vol. 2. pp. 749–752.

Snyder D, Garcia-Romero D, Sell G, Povey D, Khudanpur S, 2018. X-vectors: Robust dnn embeddings for speaker recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5329–5333.
Tachioka Y, Narita T, Ishii J, 2014. Speech recognition performance estimation for clipped speech based on objective measures. Acoust. Sci. Technol. 35 (6), 324–326.

Thomas IB, Niederjohn RJ, 1970. The intelligibility of filtered-clipped speech in noise. J. Audio Eng. Soc. 18 (3), 299–303.

Vincent E, Gribonval R, Févotte C, 2006. Performance measurement in blind audio source separation. IEEE Trans. Audio, Speech Language Process. 14 (4), 1462–1469.

Yu C, Hansen JH, Oard DW, 2014. Houston, we have a solution’: A case study of the analysis of astronaut speech during nasa apollo 11 for long-term speaker modeling. In: INTERSPEECH.
Fig. 1. Illustration of the difference using Clip-by-Amplitude and Clip-by-Sample over time definition.
Fig. 2.
Mapping of clipping rates using the Clip-by-Amplitude definition and Clip-by-Sample over time definition.
Fig. 3.
Clipping overview: comparison of original clean and clipped speech that includes waveform, spectrogram, and MFCC feature vectors for three phonemes (/ei/, /n/, /sh/). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
Fig. 4.
SID DET curve results: clean trained models; 5 test data configurations — clean, 0.5%, 1%, 5%, 10% clipped test data.
Fig. 5.
SID DET curve results using four levels of clipped training data. The test data is clean.
Fig. 6. SID DET curve results supplementing scarce training data with degraded/clipped audio. The test data is clean.
Fig. 7. SID DET curve results when removing clipped sample sequences. The test data is clean.
Fig. 8.
SID DET curve results using i-Vector and x-Vector with PLDA on NIST SRE-08 corpus: performance for 0%, 0.5%, 1%, 5%, 10% clipping distortion on the test data.

(a) i-Vector PLDA results
(b) x-Vector PLDA results
Fig. 9.
ClipDaT Clipping Detection Algorithm: Block diagram overview.
Fig. 10.
Visual demonstration of alternative waveform sequences of clipping encountered during ClipDaT development: clipped speech is not always “flat”.

Speech Commun. Author manuscript; available in PMC 2022 June 30.
Fig. 11.
Graphical representation of amount and distribution of clipping within the Pan-Arabic speech corpus.
Fig. 12.
Graphical representation of amount and distribution of clipping within the NoTel speech corpus.
Fig. 13.
Graphical representation of amount and distribution of clipping within the SUSAS corpus.
Fig. 14.
Graphical representation of amount and distribution of clipping within the Switchboard 2 Phase 2 corpus.
Fig. 15.
Graphical representation of amount and distribution of clipping within the SPINE 1 and 2 corpus.
Fig. 16.
Graphical representation of amount and distribution of clipping within the RATS corpus.
Table 1

Comparison of four speech quality measures for original clean speech, and different percentages (0.5%–10.0%) of waveform clipping on TIMIT data.

| Clipping rate (%) | Quality measure | NIST STNR | WADA SNR | PESQ | BSS Eval (SAR) |
|-------------------|-----------------|-----------|----------|------|----------------|
| 0                 |                 | 49.6      | 80.7     | –    | –              |
| 0.5               |                 | 48.8      | 83.8     | 3.9  | 17.9           |
| 1                 |                 | 48.2      | 83.1     | 3.7  | 15.7           |
| 5                 |                 | 44.6      | 77.7     | 3.1  | 10.4           |
| 10                |                 | 40.8      | 71.5     | 2.7  | 8.3            |
### Table 2
Comparison of speech quality measures and subjective human perception of different amounts of clipped audio.

| Clipping rate (%) | Quality assessment |
|-------------------|--------------------|
|                   | PESQ   | MOS   |
| 0                 | –      | 4.6   |
| 0.5               | 3.9    | –     |
| 1                 | 3.7    | 4.2   |
| 5                 | 3.1    | 4.0   |
| 10                | 2.7    | 3.8   |
| 15                | 2.1    | 3.0   |
Table 3
SID EER results: clean trained models; 5 test data configurations — clean, 0.5%, 1%, 5%, 10% clipped test data.

| Test data     | EER (%) |
|---------------|---------|
| Clean         | 3.20    |
| 0.5% clipped  | 4.13    |
| 1% clipped    | 5.02    |
| 5% clipped    | 9.69    |
| 10% clipped   | 14.61   |
Table 4
SID EER results using four combinations of clipped and unclipped training data. The test data is clean.

| Train data       | EER (%) |
|------------------|---------|
| 8 Clean sentences| 3.20    |
| 6 Clean, 2 Clipped | 3.95  |
| 4 Clean, 4 Clipped | 5.01  |
| 2 Clean, 6 Clipped | 7.63  |
Table 5
SID EER results supplementing scarce training data with degraded/clipped audio. The test data is clean.

| Train data           | EER (%) |
|----------------------|---------|
| 2 Clean Sentences    | 7.76    |
| 2 Clean, 2 Clipped   | 7.11    |
| 2 Clean, 4 Clipped   | 6.91    |
| 2 Clean, 6 Clipped   | 7.63    |
Table 6
SID EER results when removing clipped sample sequences. The test data is clean.

| Train data  | Clip removal | EER (%) |
|-------------|--------------|---------|
| 2 Clean, 6 Clipped | No           | 7.63    |
|              | Yes          | 6.80    |
| 4 Clean, 4 Clipped  | No           | 5.01    |
|              | Yes          | 4.61    |
| 6 Clean, 2 Clipped  | No           | 3.95    |
|              | Yes          | 3.95    |
Table 7

SID EER results using i-Vector and x-Vector PLDA methods on NIST SRE-08 with different test data configurations: clean, 0.5%, 1%, 5%, 10% clipped test data.

| Test data    | i-Vector (%) | x-Vector (%) |
|--------------|--------------|--------------|
| Clean        | 4.82         | 2.19         |
| 0.5% clipped | 4.39         | 2.63         |
| 1% clipped   | 4.82         | 2.63         |
| 5% clipped   | 6.58         | 3.95         |
| 10% clipped  | 8.33         | 7.02         |
| Corpus            | Total duration (h) | Language                        | Clean files (%) | Clipped files (%) | Clipped files (h) | Clipped data (s) |
|-------------------|--------------------|--------------------------------|-----------------|-------------------|-------------------|------------------|
| Pan-Arabic        | 22.22              | Arabic (5 regional dialects)    | 81.46           | 18.54             | 4.12              | 5.64             |
| Noisy Telephone   | 166.25             | English                         | 72.91           | 27.09             | 45.03             | 12.82            |
| SUSAS             | 3.88               | English                         | 77.45           | 22.55             | 0.88              | 29.22            |
| Switchboard 2 P2  | 745.22             | English                         | 88.56           | 11.44             | 85.28             | 17.59            |
| SPINE 1&2         | 81.31              | English                         | 91.47           | 8.53              | 6.94              | 0.88             |
| RATS E63          | 1825.2             | Arabic, Farsi, Dari, Urdu, Pashtu| 69.02           | 30.98             | 565.53            | 15045.01         |