NOISE-ROBUST DETECTION OF PEAK-CLIPPING IN DECODED SPEECH

James Eaton and Patrick A. Naylor

Department of Electrical and Electronic Engineering, Imperial College, London, UK

{j.eaton11, p.naylor}@imperial.ac.uk

ABSTRACT

Clipping is a commonplace problem in voice telecommunications and detection of clipping is useful in a range of speech processing applications. We analyse and evaluate the performance of three previously presented algorithms for clipping detection in decoded speech in high levels of ambient noise. We identify a baseline method which is well known for clipping detection, determine experimentally the optimized operation parameter for the baseline approach, and use this in our experiments. Our results indicate that the new algorithms outperform the baseline except at extreme levels of clipping and negative signal-to-noise ratios.

Index Terms— speech enhancement, clipping detection, signal recovery

1. INTRODUCTION

Peak-clipping occurs when the amplitude of an input signal to an audio device has exceeded the available dynamic range of the device. Peak-clipping in speech is undesirable because it reduces the subjective quality [1], may result in a loss of intelligibility, and affects the performance of subsequent speech processing such as Automatic Speech Recognition (ASR). Whilst mobile handset manufacturers are making efforts to cope with high amplitude signals such as Nokia's High Amplitude Audio Capture (HAAC) technology [2], the use of low cost electronic components, increasing miniaturisation, and high levels of ambient noise [3] contribute to a high likelihood of the occurrence of distortion. Since distortion-limiting microphone technology is not yet widespread, there remains a need to be able to detect, and where possible correct, the effects of peak-clipping in noisy speech after the signal has been processed by one or more codecs. The scenario considered here uses 'perceptual codecs' rather than waveform coders. We use 'perceptual codec' to refer to codecs optimized for sound quality. In contrast to waveform coders, perceptual codecs do not normally preserve waveform shape [4].

In [5] we presented three methods for detection of peak-clipping and estimation of original peak signal level of a coded-decoded speech signal in noise-free conditions. We now study the performance of these algorithms in noise compared to a baseline.

For the baseline, an established method [6] for detecting clipped samples in a peak-clipped signal has been selected. This baseline method considers a signal $s(n)$ of length N containing clipped samples. The set of indices $\mathbf{c}$ at which $s(n)$ has clipped samples is defined as:

$$\mathbf{c} = \{i : 0 \leq i < N \text{ and } (x(i) > \mu^{+} \text{or } x(i) < \mu^{-})\} \quad (1)$$

where $\mu_{+} = (1-\epsilon)\ \max\{s(n)\}$ and $\mu_{-} = (1-\epsilon)\ \min\{s(n)\}$ for some tolerance ϵ. Clipping level, λ, is defined as the reciprocal of Overdrive Factor (ODF), Λ. It can be expressed relative to the full dynamic range of the signal as $\lambda = -20\log_{10}(\Lambda)$ dBFS, where Λ is the factor by which the signal has been multiplied before being clipped at the peak absolute amplitude of the original signal.

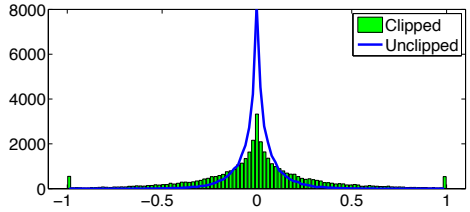

Fig. 1. *Amplitude histograms for unclipped and clipped gamma distributed signals with shape* $= 0.5$*. The clipped signal is clipped at* 30% *of the peak maximum input signal level and then amplified by* $3.3\dot{3}$

The contribution of this paper is to extend the **prior work** in [5] to investigate the performance of the Iterated Logarithm Amplitude Histogram (ILAH), Least Squares Residuals (LSR) and Least Squares Residuals Iterated Logarithm Amplitude Histogram (LILAH) algorithms in a range of noise levels and noise types alongside the baseline of (1) with ϵ optimized to noisy speech followed by coding and decoding, ϵ_o. The remainder of this paper is organised as follows: In Section 2, the peak-clipping detection algorithms first described in [5] are reviewed. In Section 3 the test approach is discussed. In Section 4 the outcomes of the tests conducted are reported, and in Section 5 the results are discussed and conclusions drawn.

2. REVIEW OF ILAH, LSR, AND LILAH

2.1. ILAH clipping detection method

The amplitude histogram of speech has been described using a gamma distribution with a shaping parameter between 0.4 and 0.5 [7, 8]. In Fig. 1, a gamma distributed signal is compared with the same signal clipped at 30 % of its peak absolute amplitude value and amplified by a factor $3.\dot{3}$, therefore a clipping level, λ, of -10.5 dBFS. The gamma distributed signal has a shaping parameter equal to 0.5 and $40,000$ samples equivalent to a 5 s speech file an 8 kHz with sample rate.

After clipping and passing through a perceptual codec such as Adaptive Multi-Rate (AMR) [10] or GSM 06.10 [11] the time domain features of clipping are obscured such that the sharp cut-off of amplitude values and peaks at the clipping level observed in the unprocessed clipped speech are very weakly preserved as discussed in [5]. In our methods, we seek a transformation of the clipped coded-decoded signal from which is it possible to determine the presence of clipping. We propose to use a transformation involving the Iterated Logarithm (IL) as this has the desirable effects of flattening the

Algorithm 1 ILAH

procedure ESTCLIPILAH($s(n)$) ▷ estimate the clipping level and clipped samples
  Normalize input signal $s(n)$
  Generate amplitude histogram $s'(i)$ from $s(n)$, with bins $x(i)$, where number of bins, $K = 25$
  $s'_l(i) \leftarrow \log s'(i)$ ▷ Generate log histogram
  Set all $s'_l(i) = -\infty$ to 0
  $s'_{ll}(i) \leftarrow \log s'_l(i)$ ▷ Generate log log histogram
  Set all $s'_{ll}(i) < 0$ to 0
  $i_{1+} \leftarrow$ Lowest i where $s'_{ll}(i) > 0$
  $i_{1-} \leftarrow$ Highest i where $s'_{ll}(i) < 0$
  $a_+x+b_+ \leftarrow$ Least Squares fit to $s'_{ll}(i_{1+}) \dots s'_{ll}(i_{max})$ along $x(i_{1+}) \dots x(i_{max})$ [9]
  $a_-x+b_- \leftarrow$ Least Squares fit to $s'_{ll}(i_{min}) \dots s'_{ll}(i_{1-})$ along $x(i_{min}) \dots x(i_{1-})$ as above
  if $a_+ < 0.005$ **then** ▷ Apply safety net for low gradients
    $a_+ \leftarrow 0.005$
  end if
  if $a_- > -0.005$ **then**
    $a_- \leftarrow -0.005$
  end if
  $\hat{x}_{c+} \leftarrow x(i_{max})$ ▷ Estimate clipped signal levels
  $\hat{x}_{c-} \leftarrow x(i_{min})$
  $\hat{x}_{o+} \leftarrow -b_+/a_+$ ▷ Estimate original signal levels
  $\hat{x}_{o-} \leftarrow -b_-/a_-$
  $\epsilon_c \leftarrow 0.21$ ▷ Determine signal status
  if $|\hat{x}_{o+} - \hat{x}_{c+}| > |\epsilon_c \hat{x}_{o+}|$ **or** $|\hat{x}_{o-} - \hat{x}_{c-}| > |\epsilon_c \hat{x}_{o-}$ **then**
    $\hat{C} \leftarrow 1$ ▷ Signal clipped
  else
    $\hat{C} \leftarrow 0$ ▷ Signal not clipped
  end if
  $\hat{x}_{o\mu} \leftarrow (-\hat{x}_{o-} + \hat{x}_{o+})/2$ ▷ Find mean original signal level
  $\hat{\lambda}_+ \leftarrow \hat{x}_{c+}/\hat{x}_{o\mu}$ ▷ Determine clipping level as ratio
  $\hat{\lambda}_- \leftarrow \hat{x}_{c-}/\hat{x}_{o\mu}$
  $c_+x + d_+ \leftarrow$ Least Squares fit of $s'_{ll}(x(i_{max} - 3))...s'_{ll}(x(i_{max}))$ ▷ Determine gradient of histogram sides
  $c_-x + d_- \leftarrow$ Least Squares fit of $s'_{ll}(x(i_{min}))...s'_{ll}(x(i_{min} - 3))$
  $\hat{P} \leftarrow 0$ ▷ Estimate codec clipping amount
  if $c_+ + 2 > a_+$ **and** $\hat{p}_{c+} < 0.742$ **then**
    $\hat{\lambda}_+ \leftarrow (1.0569 \times 2 \times (a_+ - c_+)/(d_+ - b_+) - i_{max})/\hat{x}_{o\mu}$
    $\hat{P} \leftarrow 1$ ▷ Perceptual codec detected
  end if
  if $c_- > a_- + 2$ **and** $\hat{p}_{c-} > -0.742$ **then**
    $\hat{\lambda}_- \leftarrow (1.0569 \times 2 \times (a_- - c_-)/(d_- - b_-) - i_{min})/\hat{x}_{o\mu}$
    $\hat{P} \leftarrow 1$ ▷ Perceptual codec not detected
  end if
  if $\hat{P}$ **then** ▷ Determine clipping detection threshold
    $t \leftarrow 0.009$
  else
    $t \leftarrow 0.001$
  end if
  $x_{t+} \leftarrow \hat{\lambda}_+ \times (1 - t) \times \hat{x}_{o\mu}$ ▷ Set clipping limit
  $x_{t-} \leftarrow \hat{\lambda}_- \times (1 - t) \times \hat{x}_{o\mu}$
  for all n **do** ▷ Determine clipped samples
    if $s(n) > x_{t+}$ **or** $s(n) < x_{t-}$ **then**
      $\hat{c}(n) \leftarrow 1$
    else
      $\hat{c}(n) \leftarrow 0$
    end if
  end for
  return $\hat{\mathbf{c}}$, Estimated original signal level, $\hat{x}_{o+}, \hat{x}_{o-}$, Estimated clipping levels, $\hat{\lambda}_+, \hat{\lambda}_-, \hat{C}, \hat{P}$
end procedure

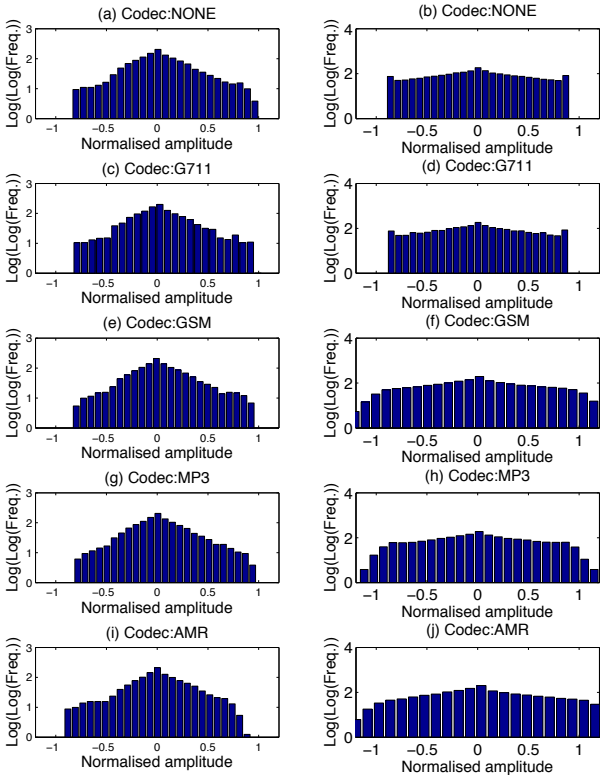

Fig. 2. *ILAHs of TIMIT file SI1027.WAV for each codec with no clipping (left hand plots) and with clipping at 30% (right hand plots)*

distribution and emphasizing the tails at either extreme. The gamma distributed signal in Fig. 1 has a large peak close to zero.

The Strong law of large numbers [12, 13] suggests that taking the logarithm of the logarithm (the IL) of a function that approaches infinity can be approximated with a first order function. The ILAH method takes the IL of a 25 point amplitude histogram ensuring that values of zero and below are removed following each iteration as illustrated in Fig. 2 (a), (c), (e), (g) and (i), transforming the distribution and revealing features that indicate clipping. Where the clipped speech has subsequently passed through a codec, the extremal values of the ILAH show a characteristic spreading so that the edges of the histogram are seen to slope outwards as Fig. 2 (f), (h) and (j). A generalised ILAH for a clipped coded-decoded speech signal is shown in Fig. 3. Estimates for the peak negative pre-clipping signal amplitudes can be obtained by fitting line (a) to the upper left side of the histogram (b) and extending this to the point where it crosses the x-axis (d) to give the estimate, and similarly with the upper right side (c). In order to prevent over-estimation of the unclipped signal level, in the case where the gradient estimate is very shallow, the gradient is limited to a suitable value. Experiments indicate that 0.005 is a suitable value.

In the case of decoded speech, the slopes (e) and (f) are due to spreading caused by the coding and decoding processes of signal amplitudes at the clipping level, λ. Thus where the sides slope

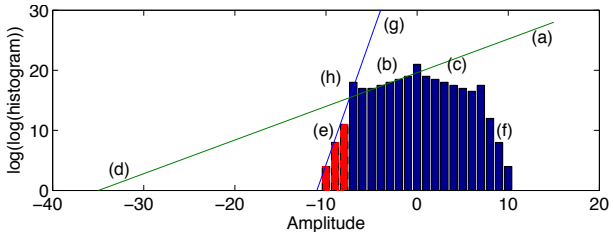

Fig. 3. *Generalised ILAH for a speech signal*

outwards, the amplitude values at the point at which each side meets each uppermost side (b) and (c) at (h) for example can be considered to be an improved estimate for the clipping level. An estimate of the amount of clipping in both an unprocessed and a coded-decoded signal can be made by estimating the gradients of sides (e) and (f) by applying a threshold to the two gradients below which the second estimate does not apply, and comparing the estimate of the peak unclipped signal level and the maximum clipped signal amplitude. Equation (1) can then be used to estimate which samples in $s(n)$ are clipped by setting ϵ equal to the clipping amount. In this way ILAH adapts to the differences in the histograms of different speakers and utterances. The ILAH method is tabulated in Algorithm 1.

Algorithm 2 LSR

procedure ESTCLIPLSR($s(n), \hat{z}(n)$) ▷ estimate the clipped samples at $\hat{z}(n)$
  Normalize input signal $s(n)$
  Generate periodogram P, of $s(n)$ using a 32-point Fast Fourier Transform (FFT), window length $= 4$ zero-padded to 32, overlap $= 75\%$
  for all $\hat{z}(n)$ frames **do**
    Fit 1st order polynomial to frequency bands in a Least Squares sense up to 3.3 kHz (bands 1 to 7)
    $r(n) \leftarrow$ (residuals of polyfit)
  end for
  Normalise r
  for all r **do**
    if $r(n) > 0.39630$ **then**
      Register of estimated clipped samples, $\hat{c}(n) \leftarrow 1$
    else
      $\hat{c}(n) \leftarrow 0$
    end if
  end for
  return $\hat{c}(n)$
end procedure

Algorithm 3 LILAH

procedure ESTCLIPLILAH($s(n)$) ▷ Estimate the clipped samples and the clipping level
  $\hat{c}(n), \hat{x}_{o+}, \hat{x}_{o-}, \hat{\lambda}_+, \hat{\lambda}_-, \hat{C}, \hat{P} \leftarrow$ ESTCLIPILAH($s(n)$)
  if $\hat{P}$ **then**
    Determine clipping zones, $\hat{z}(n)$
    $\hat{c}(n) \leftarrow$ ESTCLIPLSR($s(n), \hat{z}(n)$)
  end if
  return $\hat{c}(n), \hat{x}_{o+}, \hat{x}_{o-}, \hat{\lambda}_+, \hat{\lambda}_-, \hat{C}, \hat{P}$
end procedure

2.2. LSR clipping detection method

Clipping introduces additional harmonics and intermodulation products [14]. Whilst passing speech through a perceptual codec limits the frequency response and itself introduces distortion, some of the spectral characteristics of clipped speech are retained [15]. Spectral roughness first discussed in [16] relates to the dissonances between the harmonics of the signal. We expect a clipped signal to exhibit a high degree of spectral roughness due to the generated harmonics. We therefore propose to detect clipping in the frequency domain by estimating spectral roughness using a novel approach.

We compute a periodogram of the signal using an FFT of length 2 ms with a Hamming window of length 0.25 ms zero-padded to 2 ms, and an overlap of 75%. We then fit a line across the frequency bins for each frame creating a vector containing the residuals of the Least Squares fit for each frame. The vector is then normalized over the entire signal. High residuals indicate spectral roughness and thus clipping, and by setting a threshold above which we assume a sample to be clipped, we create a vector $\mathbf{c}$ indicating the presence of clipped samples. The optimum threshold is determined by finding the intersection of the False Positive Rate (FPR) and False Negative Rate (FNR) curves [17] for the algorithm using a suitable training corpus, where FPR is the ratio of samples incorrectly identified as clipped to the total number of unclipped samples and FNR is the ratio of samples incorrectly identified as unclipped to the total number of clipped samples. This optimum threshold was found to be 0.3963 using the TIMIT [18] training corpus. Whilst accuracy is better than with ILAH, the cost of computing a Least Squares fit for every frame is high, and no estimate for the clipping level or unclipped signal level is obtained. We refer to this method as LSR and it is presented in Algorithm 2.

2.3. Combining LSR and ILAH methods

LSR and ILAH can be combined to produce an accurate clipping detector that also provides an estimate of the clipping level and peak unclipped signal level. In order to exploit the accuracy of LSR without the computational cost penalty, LSR is only computed in regions of the signal where ILAH indicates clipping may be present. We refer to such regions as clipping zones. This is achieved by taking the results of ILAH and applying LSR only on these regions. Clipped samples less than 20 ms apart comprise a clipping zone. We refer to this method as LILAH as presented in Algorithm 3.

3. TEST METHODOLOGY

We have evaluated all methods at 10 clipping levels and with four codecs using the Receiver Operating Characteristic (ROC) [17] to analyse the results in comparison with a suitable baseline. The same 24 male and 24 female speech files were randomly selected from TIMIT. Babble, white and volvo noises from NOISEX-92 [19] were mixed with each speech file at infinity, 15, 10, 5, 0, and -5 dB Signal-to-Noise Ratio (SNR). The codecs used were G.711 [20], GSM 06.10 [11], MP3 [21], and AMR [10] at 4.75 kbps.

We determined experimentally ϵ_o, the optimum ϵ for the baseline [6] in (1) across all codecs, noises and noise levels in our evaluation. This was achieved by finding the intersection of the FPR and FNR curves for the algorithm [17], where FPR is the ratio of samples incorrectly identified as clipped to the total number of unclipped samples and FNR is the ratio of samples incorrectly identified as unclipped to the total number of clipped samples. The results shown in Fig. 4 indicate $\epsilon_o = 0.71$, meaning that all samples greater than

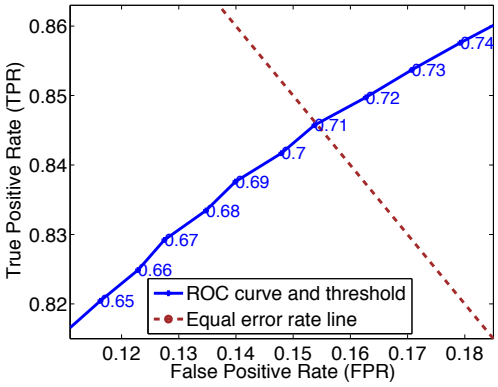

Fig. 4. *ROC curve for identification of optimum threshold for clipping detector in [6] in noisy clipped decoded and unprocessed speech*

29% of full scale amplitude are classified as clipped by the baseline method.

4. RESULTS

Accuracy (ACC) and F1 Score (F1) [17] results for unencoded and decoded speech using GSM 06.10 with no noise, and with babble noise at 10 and 0 dB SNR are shown in Fig. 5 (a) to (l) respectively for all algorithms under test. The optimized baseline performs similarly under all conditions, albeit with a low detection accuracy. Performance improves at low clipping levels since under these conditions most samples will be clipped, and ϵ_o includes most samples.

As expected the averaging nature of the histogram-based approach of ILAH provides robustness to noise. Performance gradually degrades from around 5 dB SNR. Other results show that the method is not robust to noise at SNRs of 0 dB and below due to noise adversely affecting the shape of the histogram. As in the noise-free case in [5], LSR and LILAH algorithms outperform ILAH since they are less dependent on the pdf of the signal for decoded speech. Other experiments showed similar results for babble noise, despite the more Laplacian pdf.

5. DISCUSSION AND CONCLUSIONS

Clipping is a common problem in voice telecommunications and detection of clipping is useful in a number of speech processing applications. The application in [2] suggests that sound pressure might exceed the capability of a handset microphone by a factor of 10, and in [22] it is shown that a 30 dB variation in speech amplitude is possible for different emotions. The ILAH, LSR, and LILAH methods perform best with clipping levels between -3 and -14 dBFS which is the relevant range in typical mobile telephony applications, where lower clipping levels than -14 dBFS may be unlikely.

Overall, LILAH exceeds the performance of the optimized baseline on uncoded speech, and provides better performance on decoded speech except at very high levels of clipping and negative SNRs. In these circumstances, the baseline crudely labels all samples of significant amplitude as clipped, and is therefore often correct. Our experiments have demonstrated that the LILAH method outperforms the optimized baseline on decoded speech in noisy conditions across a wide range of clipping levels, noises, noise levels and codecs in the majority of conditions down to 5 dB.

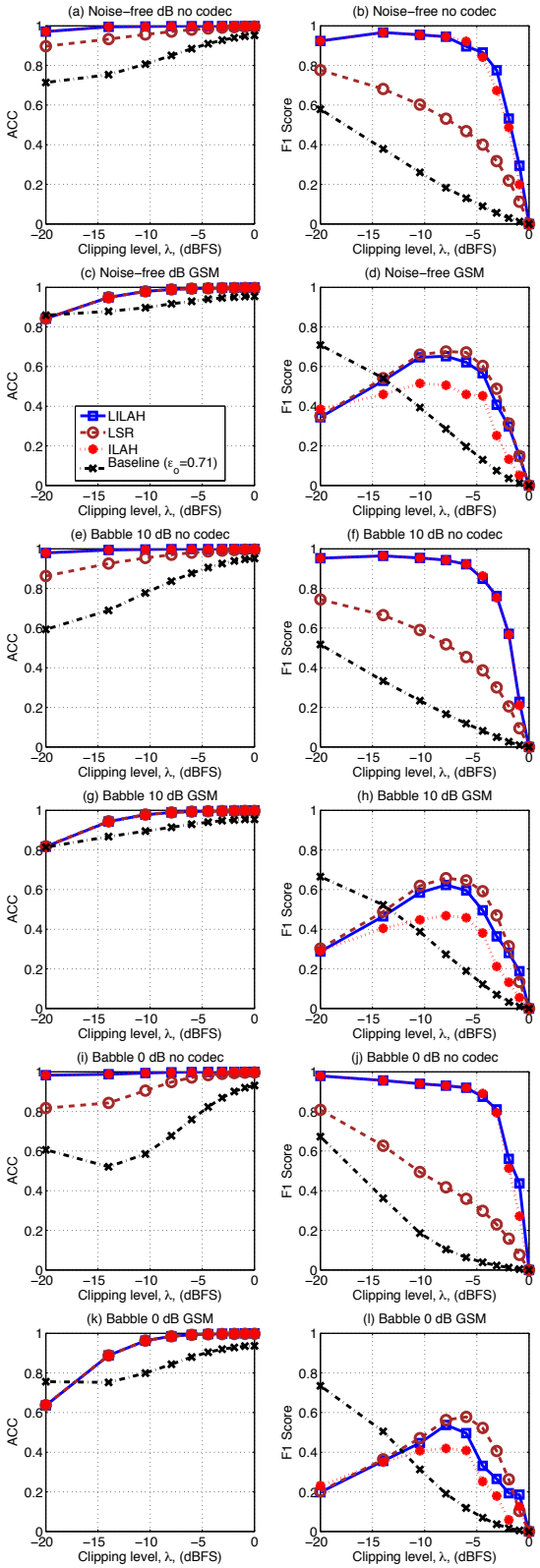

Fig. 5. *ACC and F1 using no noise and babble noise at* 10 *and* 0 dB *SNR for no codec and GSM* 06.10 *codecs at clipping levels of* 0.1 *to* 1.0

6. REFERENCES

[1] J. Gruber and L. Strawczynski, "Subjective effects of variable delay and speech clipping in dynamically managed voice systems," *IEEE Trans. Commun.*, vol. 33, no. 8, pp. 305–307, Jan. 1985.

[2] A. Koski, "Rich recording technology," Technical overall description, Nokia Product Engineering, 2012.

[3] Patrick A. Naylor and Nikolay D. Gaubitch, "Acoustic signal processing in noise: It's not getting any quieter," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, 2012, pp. 1–6.

[4] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*, Wiley, 2006.

[5] J. Eaton and P. A. Naylor, "Detection of clipping in perceptually encoded speech signals," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Marrakech, Morocco, Sept. 2013, IEEE, pp. 1–5.

[6] L. Atlas and C. Pascal Clark, "Clipped-waveform repair in acoustic signals using generalized linear prediction," U.S. Patent No. 8126578, Feb. 2012.

[7] M. Paez and T. Glisson, "Minimum mean-squared-error quantization in speech PCM and DPCM systems," *IEEE Trans. Commun.*, vol. 20, no. 2, pp. 225–230, Apr. 1972.

[8] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1978.

[9] J. F. Kenney and E. S. Keeping, *Mathematics of Statistics, Pt. 1*, chapter 15, pp. 252–285, Van Nostrand, Princeton, NJ, 3rd edition, 1962.

[10] ETSI, "Adaptive multi-rate (AMR) speech transcoding," 1998.

[11] ETSI, "GSM 06.10: European digital cellular telecommunications system (Phase 2); Full rate speech transcoding," Sept. 1994.

[12] W. Feller, "The general form of the so-called law of the iterated logarithm," *Trans. Amer. Math. Soc.*, vol. 54, pp. 373–402, 1943.

[13] P. K. Pathak and C. Qualls, "A law of iterated logarithm for stationary gaussian processes," *American Mathematical Society*, vol. 181, 1973.

[14] F. Vilbig, "An analysis of clipped speech," *J. Acoust. Soc. Am.*, vol. 27, no. 1, pp. 207–207, 1955.

[15] A. Gallardo-Antolin, F. Diaz de Maria, and F. Valverde-Albacete, "Avoiding distortions due to speech coding and transmission errors in GSM ASR tasks," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 1999, vol. 1, pp. 277–280.

[16] H. L. F. Helmholtz, *On the Sensations of Tone as a Physiological Basis for the Theory of Music*, Dover Publications, Inc., 2nd edition, 1885 (1954).

[17] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Lett.*, vol. 27, pp. 861–874, 2006.

[18] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," Technical report, National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, Dec. 1988.

[19] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 3, no. 3, pp. 247–251, July 1993.

[20] ITU-T, "Pulse Code Modulation (PCM) of voice frequencies," Nov. 1998.

[21] ISO, "ISO/IEC 11172-3:1993 Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s – Part 3: Audio," 1995.

[22] H. R. Pfitzinger and C. Kaernbach, "Amplitude and amplitude variation of emotional speech," in *Proc. Interspeech Conf.*, Brisbane, Australia, Sept. 2008, vol. 1, pp. 1036–1039.