EUSIPCO 2013 1569741837

DETECTION OF CLIPPING IN CODED SPEECH SIGNALS

James Eaton and Patrick A. Naylor

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

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

ABSTRACT

In order to exploit the full dynamic range of communications and recording equipment, and to minimise the effects of noise and interference, input gain to a recording device is typically set as high as possible. This often leads to the signal exceeding the input limit of the equipment resulting in clipping. Communications devices typically rely on codecs such as GSM 06.10 to compress voice signals into lower bitrates. Although detecting clipping in a hard-clipped speech signal is straightforward due to the characteristic flattening of the peaks of the waveform, this is not the case for speech that has subsequently passed through a codec. We describe a novel clipping detection algorithm based on amplitude histogram analysis and least squares residuals which can estimate the clipped samples and the original signal level in speech even after the clipped speech has been perceptually coded.

Index Terms: speech enhancement, clipping detection, signal recovery

1. INTRODUCTION

Clipping is caused when the input signal to a recording device has exceeded the available dynamic range of the device. It is generally undesirable and significantly affects the subjective quality of speech [1]. Detection of clipping is therefore important in maintaining speech quality, and is employed in restoration, denoising and de-clicking applications. Detection of clipping is straightforward in raw clipped speech due to the characteristic flattening of the peaks of the waveform at the limits of the input dynamic range. We define the clipping level to be the fraction of the unclipped peak absolute signal amplitude to which a sample exceeding this value will be limited. For example in a signal clipped with a clipping level of 0.5, any input signal exceeding 50% of peak absolute amplitude will be limited to 50 % of peak absolute amplitude, and a clipping level of 1.0 will therefore leave the signal unchanged. We define Overdrive Factor (ODF) as the reciprocal of clipping level. An established method [2] for detecting clipped samples in a clipped signal considers a signal $x(n)$ of length N containing clipped samples. The set of indices $\mathbf{c}$ at which $x(n)$ 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

$$\mu_{+} = (1-\epsilon)\ \max\{x(n)\} \quad \text{and} \quad \mu_{-} = (1-\epsilon)\ \min\{x(n)\}$$

for some tolerance ϵ such as 0.01. Another clipping detection method described in [3] exploits the properties of the amplitude histogram of the signal to identify which samples are clipped. These methods work well when applied directly to the clipped signal.

1.1. Effect of a perceptual codec

In this work we define a perceptual codec to mean a lossy codec optimised for speech perception. Perceptual codecs such as GSM 06.10 which remove information not typically perceived by the listener do not in general preserve signal phase [4]. This affects the flattened peaks of a clipped signal resulting in an amplitude histogram resembling that of an unclipped signal. This greatly reduces the accuracy of clipping detection for coded speech.

Fig. 1 shows the time domain waveforms and their amplitude histograms for TIMIT [5] utterance SX12.WAV directly and through different codecs. Plots (a) and (b) are for the unprocessed utterance, whilst plots (c) to (h) show the utterance after passing through Pulse-Code Modulation (PCM) of Voice Frequencies (G.711), GSM 06.10, Moving Picture Experts Group (MPEG)-2 Audio Layer III (MP3), and Adaptive Multi-Rate (AMR) at 4.75 kbps respectively. Fig. 2 shows the same utterances clipped with a clipping level of 0.5 prior to passing through each codec. In Figs. 2 (a) to (d), the characteristic flattening of the waveform peaks and corresponding spikes in the amplitude histogram are clearly visible when compared with Figs. 1 (a) to (d). However, with a perceptual codec, the waveform and amplitude histograms for the clipped utterance are similar to the unclipped utterance (Fig. 2 (e) to (j) and Fig. 1 (e) to (j)).

Established clipping detectors and restoration algorithms such as those presented in [6, 7] rely on these time domain features and may fail when presented with a post-codec speech signal. In [8] a spectral transformation of each frequency band using a model spectral envelope is proposed. This method may work on post-codec speech if trained on post-codec speech and used with a codec detector, but is outside of the scope of this paper.

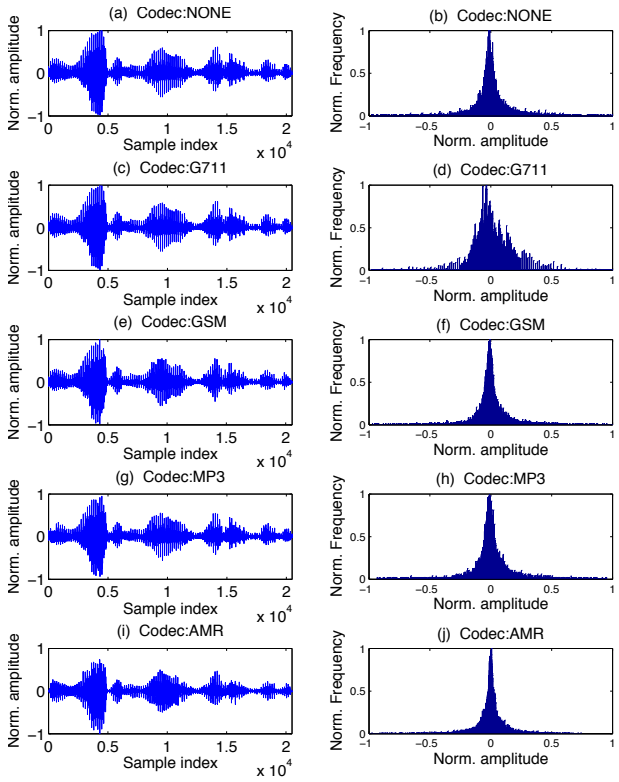

Fig. 1. *Waveforms and amplitude histograms for the unclipped signals.*

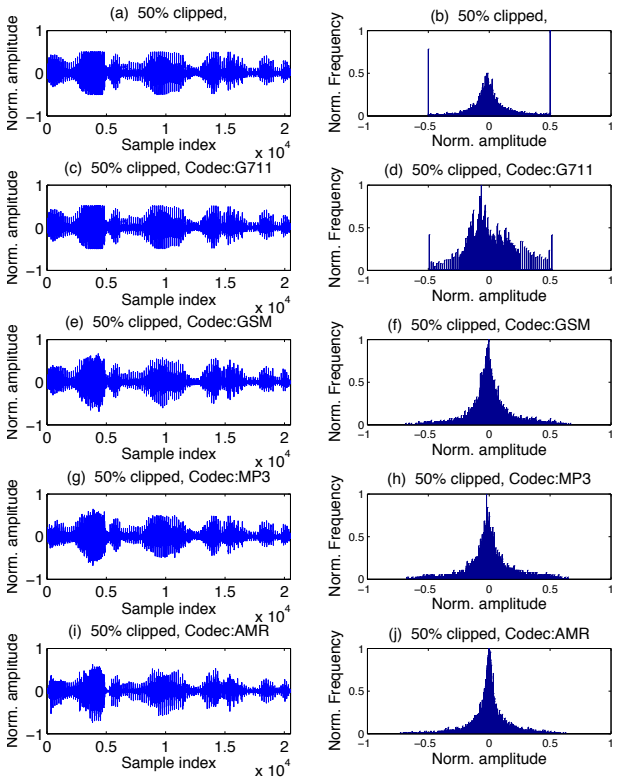

Fig. 2. *Waveforms and amplitude histograms for the clipped signals*

The key contributions of this paper are to: (1) propose a non-intrusive clipping detector for speech that may have passed through a codec employing the Least Squares Residuals Iterated Logarithm Amplitude Histogram (LILAH) method; (2) show how this is robust to perceptual codecs; and to show a comparison of the results of the proposed methods with a clipping detector from the literature [2].

2. PROPOSED METHOD

We now introduce the novel Iterated Logarithm Amplitude Histogram (ILAH) method to detect clipping and unclipped signal level, and the Least Squares Residuals (LSR) method by frame in the frequency domain to reduce estimation errors. We further present LILAH which uses the ILAH method to reduce the computational complexity of LSR.

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 [9]. After clipping and passing through a perceptual codec such as GSM 06.10 the time domain features of clipping are obscured as discussed in Sec. 1.1. The Strong law of large numbers suggests that taking the logarithm of the logarithm (the Iterated Logarithm (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. 3 (a), (c), (e), (g) and (i), transforming the distribution recovering features that indicate clipping. Where the clipped speech has subsequently passed through a perceptual 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. 3 (f), (h) and (j).

A generalised ILAH for a clipped post-codec speech signal is shown in Fig. 4. An estimate for the peak negative unclipped signal amplitude 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 such as 0.005.

In the post-codec case, the sloping sides (e) and (f) represent the spread of signal levels caused by the perceptual codec. Thus where the sides slope 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

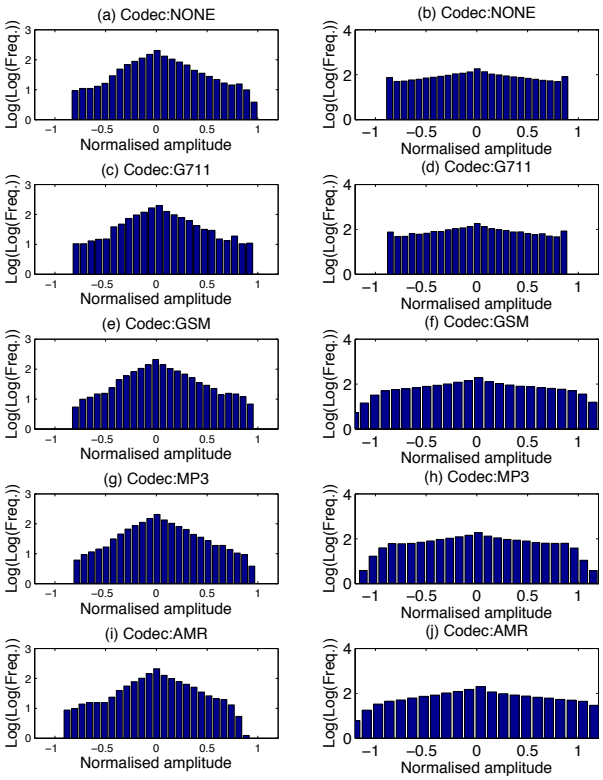

Fig. 3. *ILAHs of TIMIT file SI1027.WAV for each codec with no clipping (left hand plots) and with clipping at 30% (right hand plots)*

amount of clipping in both an unprocessed and a post-codec 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. The clipping amount and Eq. (1) can then be used to estimate which samples in $x(n)$ are clipped. We refer to this method as the Iterated Logarithm Amplitude Histogram (ILAH) method.

2.2. LSR clipping detection method

When speech is clipped, new frequencies are introduced in the form of additional harmonics and intermodulation products [10]. 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 [11]. Therefore by estimating spectral roughness we can additionally detect clipping using frequency domain processing. To achieve this, we compute a periodogram of the signal using an Fast Fourier Transform (FFT) of length 32, a Hamming window of length 4 and overlap of 75%, and then fit a line across the frequency bins for each frame in a Least Squares (LS) sense. Next we store the residuals by sample and

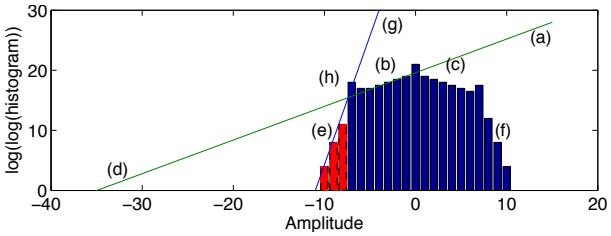

Fig. 4. *Generalised ILAH for a speech signal*

then normalise 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 can create a vector 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 for the algorithm [12], 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. Whilst accuracy is better than with ILAH, the cost of computing a LS fit for every sample is high, and no estimate for the clipping level or unclipped signal level is obtained.

2.3. Combining LSR and ILAH methods

We propose to combine LSR and ILAH to produce an accurate clipping detector that also provides an estimate of the clipping level and peak unclipped signal level. Here we only compute the LSR where there is an indication of clipping from ILAH reducing computational complexity. This is achieved by taking the results of the ILAH clipping test, and establishing clipping zones in the time domain where LSR will operate using a 20 ms rolling window: if clipped samples are less than 20 ms apart they comprise a zone. LSRs are only computed within zones, and samples outside of the zones are assumed to be unclipped. The computational complexity is therefore dependent on the estimated clipping level. We refer to this method of clipping detection as the LILAH.

3. TEST METHODOLOGY

The approach to testing was to compare all methods at 10 clipping levels and with the four codecs in Fig. 1 using a large number of speech files, and to use the Receiver Operating Characteristic (ROC) [12] to analyse the results. A set of 24 male and 24 female speech signals were randomly selected from the utterances of the TIMIT [5] test dataset and clipped at ten clipping levels, 0.1 to 1.0 in steps of 0.1. A ground truth binary vector $\mathbf{c}$ of clipped samples was established for each utterance using (1). The clipped speech was then passed either directly or via one of four codecs: G.711, GSM 06.10,

AMR narrowband coding at 4.75 kbps, and MP3 at 128 kbps with output sample rate 8 kHz before being passed to each algorithm. We employed as a baseline the method described in (1) with $\epsilon = 0.01$ and conducted the test on the proposed ILAH, LSR and LILAH methods. We also tested the baseline ($\epsilon = 0.22$) because this was found through Equal Error Rate (EER) analysis to work well with GSM 06.10. We refer to this as the optimized baseline. Signals were time aligned to compensate for any time shifts introduced by the codecs.

We used the measures Accuracy (ACC) and F1 Score from ROC [12] to compare detection performance of each algorithm. ACC is a measure of how accurately the algorithms identify clipped and unclipped samples as a percentage of the total number of samples. F1 Score is a measure of the correlation between the vector **c** for each algorithm and the ground truth and is a guide to overall performance. The measures are computed as follows:

$$\text{ACC} = (TP + TN)/(P + N) \quad (2)$$

$$\text{F1 Score} = 2TP/(2TP + FP + FN) \quad (3)$$

where: TP is the number of True Positives, (samples correctly identified as clipped); TN is the number of True Negatives (samples correctly identified as unclipped); FP is the number of False Positives (samples incorrectly identified as clipped); FN is the number of False Negatives (samples incorrectly identified as unclipped); P is the total number of samples identified as clipped, both correctly and incorrectly, and N is the total number of samples identified as unclipped, both correctly and incorrectly.

Computational complexity was compared using estimated Real-Time Factor (RTF) for each algorithm. Mean elapsed processing time using the Matlab *tic* and *toc* functions for each call on a 2.3 GHz Intel i5 Core processor with 4 GB 1.333 GHz DDR3 SDRAM was divided by the mean speech file duration over all tests to give RTF. All implementations were in Matlab.

4. RESULTS AND DISCUSSION

ACC and F1 Scores averaged over 48 TIMIT files for each codec and clipping level are shown in Fig. 5 for the methods: baseline, the optimized baseline, the proposed ILAH, LSR, and LILAH. The ACC of the baseline where no codec is used exceeds the proposed methods at clipping levels of 0.3 and below as shown in Fig. 5 (a) and (c). With a perceptual codec however, the ACC of the proposed algorithms exceeds the performance of both the baseline and the optimized baseline. ILAH performs better at lower clipping levels (higher ODF) because it adapts to each speaker and utterance, whilst a fixed threshold does not. All algorithms generate few FPs with the exception of the optimized baseline with no codec and G.711 because with a codec at least some of the positives are correct, but with no codec most of the positives are incorrect. For the

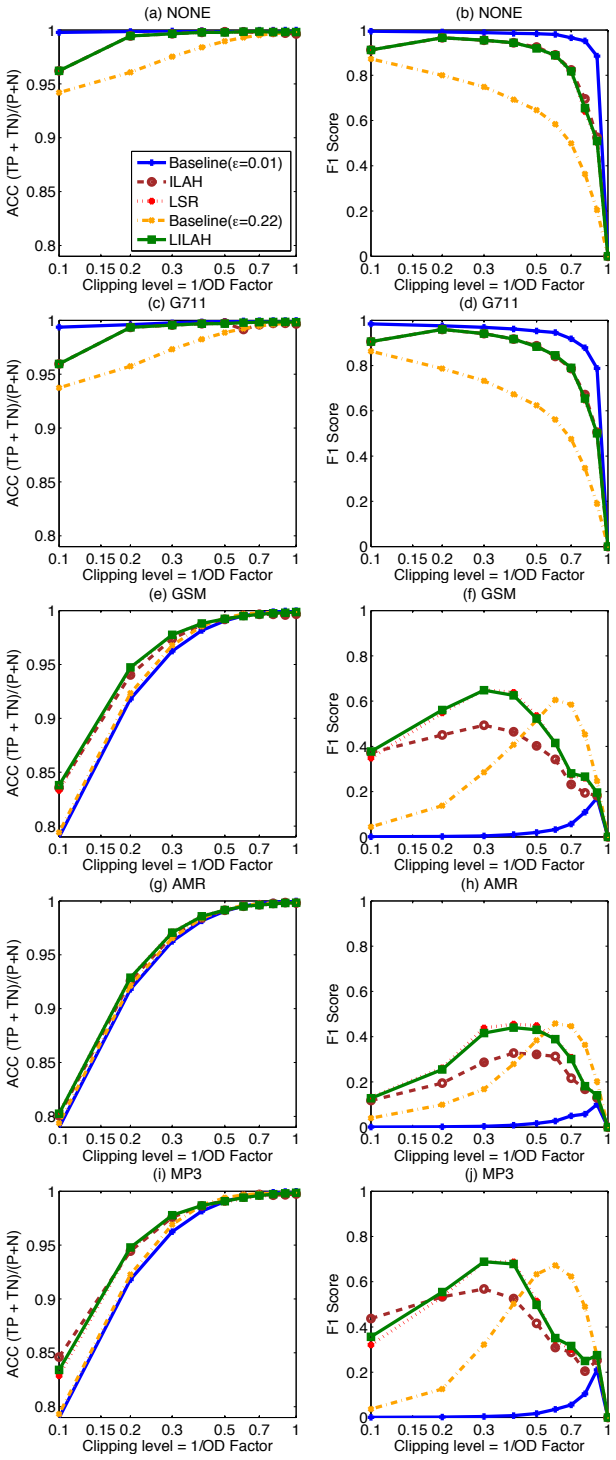

Fig. 5. *ACC and F1 by codec by algorithm and by clipping level*

proposed algorithms, more FPs are identified but more TPs are identified in the presence of a codec also giving a greater ACC score under these conditions.

The F1 Score of the proposed methods where no codec

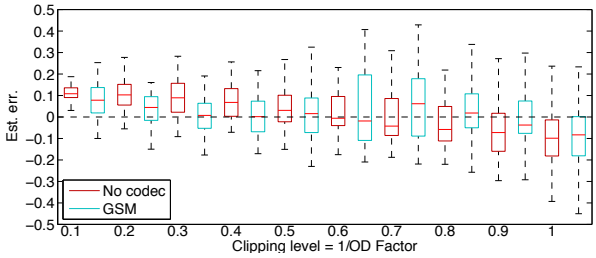

Fig. 6. *Unclipped peak absolute signal level estimation accuracy for no codec and GSM* 06.10 *by clipping level*

Table 1. *Mean Real-Time Factor by algorithm.*

Baseline($\epsilon = 0.01$)	ILAH	LSR	LILAH
0.00168	0.0075	17.7	3.65

is used is similar to the baseline at low clipping levels (high ODFs), but when a perceptual codec is used the F1 Score far exceeds the baseline as shown in Fig. 5, because the proposed methods correctly identify many more clipped samples. The optimized baseline performs moderately well in a codec at high clipping levels (low ODF) where the codec, speaker and utterance have little impact, but it fails at low clipping levels since ϵ needs to vary with each speaker and utterance to get good results whereas ILAH has an adaptive threshold, and LSR uses frequency domain features independent of threshold. The improved F1 Score of LSR over ILAH, and that combining the two methods into LILAH overall improves the F1 Score is shown in Fig. 5 (f), (h) and (j).

These results show that the LILAH algorithm shows some robustness to the codec used providing a non-intrusive estimate of the clipped samples in speech with comparable performance to the baseline in uncoded signals, and better than the optimized baseline for perceptually coded signals without prior knowledge of whether a codec has been used.

Using ILAH as a pre-processor for the LSR method results in substantially reduced RTF over the test range of clipping levels as shown in Table 1. Only the baseline RTF is shown since ϵ does not significantly affect RTF.

A useful feature of the ILAH method is that the peak absolute unclipped signal level is estimated as discussed in Sec. 2.1 which may find application in restoration algorithms. Estimation results for unprocessed and GSM 06.10 coded clipped signals are shown in Fig. 6.

5. CONCLUSIONS

We have proposed a novel method LILAH for detecting clipping in speech that shows robustness to the speaker, clipping level, and codec applied, and provides an estimate of the original signal level. Our results show that it outperforms the baseline case at detecting the clipped samples regardless of prior knowledge of the encoding used in the original signal.

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] L. Atlas and C. P. Clark, "Clipped-waveform repair in acoustic signals using generalized linear prediction," US Patent US8 126 578, Feb., 2012.

[3] T. Otani, M. Tanaka, Y. Ota, and S. Ito, "Clipping detection device and method," U.S. Patent 20 100 030 555, Feb. 4, 2010.

[4] J. M. Huerta and R. M. Stern, "Distortion-class modeling for robust speech recognition under GSM RPE-LTP coding," *Speech Communication*, vol. 34, no. 1-2, pp. 213–225, 2001.

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

[6] S. J. Godsill, P. J. Wolfe, and W. N. Fong, "Statistical model-based approaches to audio restoration and analysis," *Journal of New Music Research*, vol. 30, no. 4, pp. 323–338, Nov. 2001.

[7] S. Miura, H. Nakajima, S. Miyabe, S. Makino, T. Yamada, and K. Nakadai, "Restoration of clipped audio signal using recursive vector projection," in *TENCON 2011 - 2011 IEEE Region 10 Conference*, Nov. 2011, pp. 394–397.

[8] M. Hayakawa, M. Morise, and T. Nishiura, "Restoring clipped speech signal based on spectral transformation of each frequency band," *J. Acoust. Soc. Am.*, vol. 131, no. 4, pp. 3444–3444, 2012.

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

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

[11] A. Gallardo-Antolin, F. D. 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)*, vol. 1, Mar. 1999, pp. 277–280.

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