Robust Parameter Estimation for Audio Declipping in Noise

Mark J. Harvilla and Richard M. Stern

Department of Electrical and Computer Engineering
Carnegie Mellon University, Pittsburgh, PA 15213 USA
mharvill@cs.cmu.edu, rms@cs.cmu.edu

Abstract

Contemporary audio declipping algorithms often ignore the possibility of the presence of additive channel noise. If and when noise is present, however, the efficacy of any declipping algorithm is critically dependent on the accuracy with which clipped portions of the signal can be detected. This paper introduces an effective technique for inferring the amplitude and percentile values of the clipping threshold, and develops a statistically-optimal classification algorithm for accurately differentiating between clipped and unclipped samples in a noisy speech signal. The overall effectiveness of the clipped sample estimation algorithm is evaluated by the degree to which automatic speech recognition performance is improved when decoding clipped speech that has been declipped with state-of-the-art declipping algorithms paired with the clipped sample estimation algorithm. Up to 35% relative improvements in word error rate have been observed. Beyond the accuracy of the developed techniques, this paper generally underscores the necessity of robust parameter estimation methods for declipping in noise.

Index Terms: Nonlinear distortion, declipping, robust speech recognition, speech enhancement, parameter estimation

1. Introduction

Audio clipping is a commonly-encountered problem in audio engineering and related fields that consists of hard limiting the peak amplitude of an audio waveform to a fixed level. Clipping often occurs in one of three ways: (1) as a result of recording an audio signal whose peak amplitude exceeds the dynamic range limitations of the A/D converter, (2) as a result of writing improperly amplitude-normalized data to a file (*e.g.*, MATLAB's `wavwrite` function requires values in the range $[-1, 1]$), or (3) deliberately, to achieve some desired perceptual characteristic (*e.g.*, as with mastering of popular music). In many cases, clipping is perceptually undesirable, causing unpleasant distortion artifacts. Clipping distortion has been shown to significantly decrease the accuracy of automatic speech recognition (ASR) systems [1]. For these reasons, among others, a variety of declipping algorithms have been developed over the years (*e.g.* [1, 2, 3, 4, 5, 6, 7, 8, 9]).

Most declipping algorithms assume that it is known which samples are actually clipped. In the complete absence of noise this information is trivial to obtain, but when noise is added after the application of clipping, it becomes much more difficult to determine the value of the clipping level and exactly which samples had been clipped. In order to use any of the various practical declipping algorithms, one must estimate accurately which samples of the audio signal had actually been clipped, if any, and whether or not noise is present. Indeed, the identification of which samples are clipped is just as vital to the usability and success of the application as is the accuracy of the actual method to restore the original values of the clipped samples.

This paper introduces novel techniques for estimating the necessary parameters associated with clipping in order to estimate which samples of a given signal are clipped. The efficacy of the methods that are introduced is evaluated both in terms of the accuracy with which they estimate the associated clipping parameters, as well as the accuracy of existing declipping algorithms when used in conjunction with the proposed methods to identify the clipped samples. The efficacy of declipping itself is evaluated indirectly in terms of word error rate (WER) of ASR.

2. Audio Clipping and Noise

A mathematical definition of clipping that will be utilized in this paper is as follows:

$$x_c[n] = \begin{cases} x[n] & \text{if } |x[n]| < \tau \\ \tau \cdot \text{sgn}(x[n]) & \text{if } |x[n]| \geq \tau \end{cases} \quad (1)$$

In Eq. (1), $x[n]$ is an unadulterated speech signal, $x_c[n]$ is a clipped speech signal, and τ is the *clipping threshold* or *clipping level*, *i.e.*, the absolute amplitude value beyond which input signal samples are lost. In this paper, the threshold value will be expressed in terms of percentiles of the absolute value of the input speech. Thus, if $\tau = \text{P}_r$, then r percent of the speech data lies in $(-\tau, +\tau)$ and $(100 - r)$ percent of the data is clipped. Computing τ in this fashion causes the effect of clipping to be independent of arbitrary scaling of the waveform, allowing for more controlled experiments.

Furthermore, this paper assumes the presence of independent additive white Gaussian channel noise, $w[n]$, which combines with the clipped signal to create the observed signal, $y[n]$, as follows:

$$y[n] = x_c[n] + w[n] \quad (2)$$

This model was adopted because it appeared to be an accurate model of speech samples from the DARPA RATS program (*e.g.* [10]). To the extent that the declipping is effective, the effects of noise that is added before the clipping distortion could be mitigated by approaches such as vector Taylor series [11] or spectral subtraction [12].

In this paper, the SNR of the observed signal, y_n, is computed with respect to the clipped signal, and thus does not account for noise associated with clipping distortion.

3. Parameter Estimation

This section develops the necessary tools for blindly inferring qualities of an incoming speech signal such that it can be ac-

curately determined whether or not the speech signal contains clipped segments, and if so, which samples of the signal comprise the clipped segments.

3.1. Clipping threshold estimation

Upon receipt of a speech signal, it must be determined whether or not clipping is present, and if so, the corresponding value of the clipping level, τ. Clipping has a marked effect on the amplitude distribution of the audio waveform and is characterized by the presence of two sharp peaks in the distribution which are symmetric about 0 and located at $\pm\tau$. The addition of independent Gaussian noise to the clipped signal according to Eq. 2 implies convolution of the Gaussian noise distribution with the clipped signal amplitude distribution, and results in three Gaussian-like lobes at 0 and $\pm\tau$.

Based on these observations, it stands to reason that information concerning the location of the peaks in the noisy clipped speech amplitude distribution can be leveraged to design a method of estimating τ. Consider a sequence of data with K peaks whose locations with respect to the independent variable are $\{k_0, k_1, k_2, ..., k_{K-1}\}$. If an algorithm to estimate these locations is applied to $y[n]$[1], an estimate of the value of τ is given by:

$$\tilde{\tau} = \frac{1}{K-1} \sum_{i=0}^{K-1} |k_i| \tag{3}$$

When clipping of speech has occurred, $K = 3$ and the individual peaks should in principle be found at $k_0 = -\tau$, $k_1 = 0$, $k_2 = \tau$; the sum in Eq. 3 is then effectively $\frac{2\tau}{2} = \tau$. If no clipping has occurred, $K = 1$, and the result diverges to ∞, which is correct (*i.e.*, if no clipping has occurred, then the clipping level is effectively infinite). Thus, this technique simultaneously performs regression to predict τ and binary classification to determine whether or not the speech has been clipped at all.

3.2. Threshold percentile estimation

As noted in Sec. 2, the clipping threshold is often expressed in terms of percentiles. As will become evident in Sec. 3.3, being able to infer not only the amplitude value of τ, but also the percentile value of τ, from an observed noisy clipped speech waveform will be a useful aid to the determination of which samples are clipped.

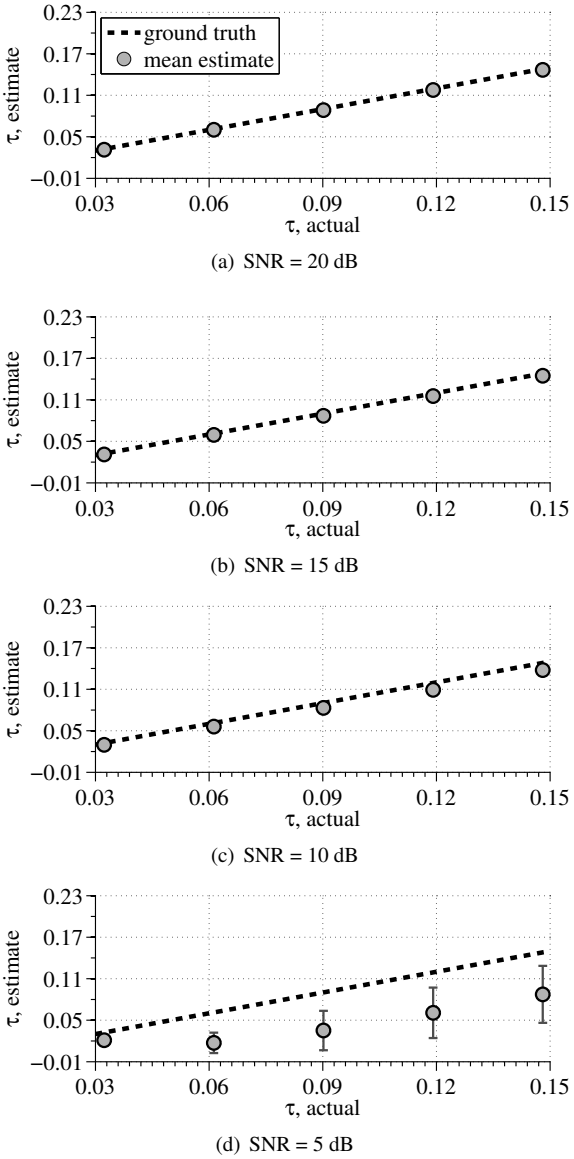
The percentile value of τ is equal to the integral (or cumulative sum) of the probability density function of the observed speech between amplitudes $-\tau$ and $+\tau$. Mathematically, where $c(x)$ is the PDF of the observed speech, and $C(x)$ is the corresponding cumulative distribution function (CDF),

$$\text{percentile value of } \tau = \int_{-\tau}^{+\tau} c(x)dx \tag{4a}$$

$$= \int_{-\infty}^{+\tau} c(x)dx - \int_{-\infty}^{-\tau} c(x)dx \tag{4b}$$

$$= C(\tau) - C(-\tau) \tag{4c}$$

[1]From basic calculus, the peaks of a signal can be found by finding the zeros of the first derivative of the signal.



(a) SNR = 20 dB

(b) SNR = 15 dB

(c) SNR = 10 dB

(d) SNR = 5 dB

Figure 1: *Results of blindly predicting τ as described in Sec. 3.1. The clipped speech is added to white Gaussian noise to achieve the indicated SNR. For a given τ value, τ is predicted over 500 independent trials of the same clipped speech added to a newly-generated white noise sequence; the markers show the sample mean of the τ predictions; the error bars extend one standard deviation above and below the mean.*

3.3. Clipped sample estimation

In the presence of noise according to Eq. 2, the identification of which samples are clipped – even given the value of τ – is not trivial. Because the addition of noise perturbs the amplitude of the signal samples, it is no longer possible to know with certainty whether the underlying speech signal's samples were clipped in a certain interval of time. We have found a probabilistic approach to be useful in making an informed decision concerning whether or not a given (series of) sample(s) is clipped.

In particular, the identification of clipped samples is a binary classification problem (*i.e.*, a sample is either clipped or not). For simplicity, it may be assumed that the probability of any given sample being clipped is a function only of its observed amplitude, $y[n]$, the signal's power, σ_y^2, the variance

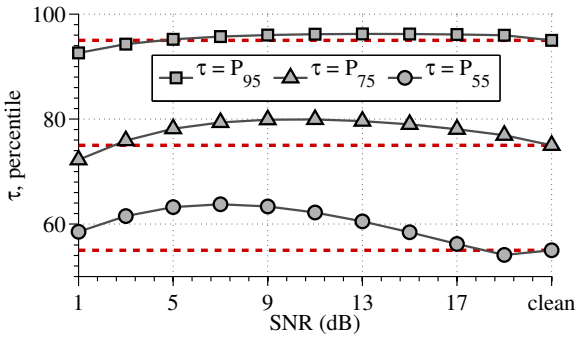


Figure 2: *Results for predicting the percentile of value of τ given its amplitude value for a particular speech signal. As in Fig. 1, the markers reflect the sample mean of 500 independent predictions, where a new white noise sequence was generated for each trial. The red dashed lines indicate the target (true) percentile.*

(power) of the white Gaussian noise, σ_w^2, and the (given) value of τ.

To begin, it would be useful to determine the conditional probability that the output of the clipping function of Eq. 1 is equal to $\pm\tau$, given the above information. Proceeding mathematically, the intention is to compute:

$$\Pr(x_c[n] = \pm\tau | y[n], \sigma_y^2, \sigma_w^2, \tau) \tag{5}$$

Using Bayes' theorem [13],

$$\Pr(x_c[n] = \pm\tau | y[n], \sigma_y^2, \sigma_w^2, \tau) = \frac{\Pr(y[n] | x_c[n] = \pm\tau, \sigma_y^2, \sigma_w^2, \tau) \Pr(x_c[n] = \pm\tau | \sigma_y^2, \sigma_w^2, \tau)}{\Pr(y[n] | \sigma_y^2, \sigma_w^2, \tau)} \tag{6}$$

The numerator can be simplified slightly by noting that the probability of $x_c[n]$ being clipped is independent of the signal and noise power, and as will be shown, the probability of $y[n]$ given that $x_c[n] = \pm\tau$ is independent of the overall signal power. Also, the denominator can be expanded, as follows:

$$\begin{aligned}\Pr(x_c[n] = \pm\tau | y[n], \sigma_y^2, \sigma_w^2, \tau) = \\ \Pr(y[n] | x_c[n] = \pm\tau, \sigma_w^2, \tau) \Pr(x_c[n] = \pm\tau | \tau) / \\ \Pr(y[n] | x_c[n] = \pm\tau, \sigma_w^2, \tau) \Pr(x_c[n] = \pm\tau | \tau) \\ + \Pr(y[n] | x_c[n] \neq \pm\tau, \sigma_y^2, \sigma_w^2, \tau) \Pr(x_c[n] \neq \pm\tau | \tau)\end{aligned} \tag{7}$$

Under the assumption of zero-mean additive white Gaussian noise (AWGN) with variance σ_w^2, the probability of the noisy signal having observed value $y[n]$ given that $x_c[n] = \pm\tau$ is:

$$\Pr(y[n] | x_c[n] = \pm\tau, \sigma_w^2, \tau) = \lim_{\epsilon \to 0} \int_{|y[n]|-\epsilon}^{|y[n]|+\epsilon} \frac{1}{\sqrt{2\pi}\sigma_w} e^{-\frac{(t-\tau)^2}{2\sigma_w^2}} dt \tag{8}$$

Moreover, the probability that a given sample $x_c[n]$ is equal to $\pm\tau$ is related to the percentile value of τ:

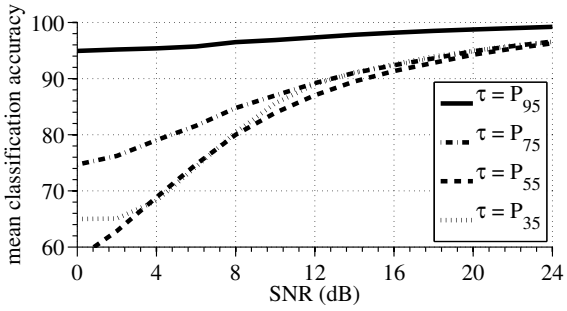


Figure 3: *Mean classification accuracy for classifying individual noisy samples as either clipped or not clipped using the rule in Eq. 13.*

$$\Pr(x_c[n] = \pm\tau | \tau) = 1 - \text{percentile value of } \tau \tag{9}$$

Furthermore,

$$\Pr(x_c[n] \neq \pm\tau | \tau) = 1 - \Pr(x_c[n] = \pm\tau | \tau) \tag{10}$$

The last term to define is the conditional probability of the observed sample, $y[n]$, given that the underlying noise-free sample is not clipped. Note that $y[n] = x_c[n] + w[n]$, where both $x_c[n]$ and $w[n]$ are random variables. As described in Sec. 3.1, the PDF of $y[n]$ would be equal to the convolution of the PDF of $x_c[n]$ with the PDF of $w[n]$. Thus, this term requires the estimation of the PDF of $x_c[n]$, which is not directly observable. To avoid the complications involved in this density estimation, it will be assumed that the conditional PDF of $y[n]$ given that $x_c[n] \neq \pm\tau$ can be modeled as a Gaussian distribution with zero-mean and variance, σ_y^2, equal to the sample variance of the observed noisy speech waveform. Therefore,

$$\Pr(y[n] | x_c[n] \neq \pm\tau, \sigma_y^2, \tau) = \lim_{\epsilon \to 0} \int_{|y[n]|-\epsilon}^{|y[n]|+\epsilon} \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{t^2}{2\sigma_y^2}} dt \tag{11}$$

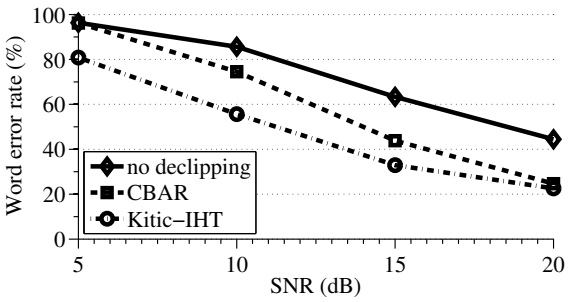
With these quantities, it is now also possible to compute the posterior probability of a sample of the noise-free signal being unclipped:

$$\begin{aligned}\Pr(x_c[n] \neq \pm\tau | y[n], \sigma_y^2, \sigma_w^2, \tau) = \\ \Pr(y[n] | x_c[n] \neq \pm\tau, \sigma_y^2, \sigma_w^2, \tau) \Pr(x_c[n] \neq \pm\tau | \tau) / \\ \Pr(y[n] | x_c[n] = \pm\tau, \sigma_w^2, \tau) \Pr(x_c[n] = \pm\tau | \tau) \\ + \Pr(y[n] | x_c[n] \neq \pm\tau, \sigma_y^2, \sigma_w^2, \tau) \Pr(x_c[n] \neq \pm\tau | \tau)\end{aligned} \tag{12}$$

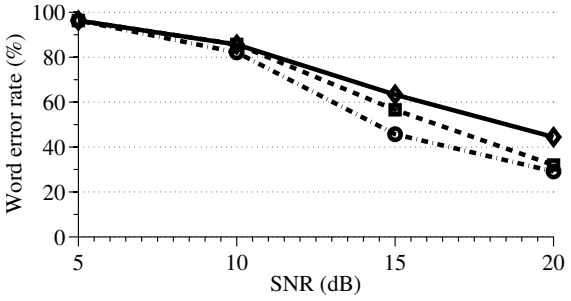
A given observed signal sample, $y[n]$, can be classified as either "clipped" or "unclipped" according to the optimal Bayesian decision threshold [14] as follows:

$$\text{class of } y[n] = \begin{cases} \text{clipped} & \text{if } \frac{\Pr(x_c[n] = \pm\tau | y[n], \sigma_y^2, \sigma_w^2, \tau)}{\Pr(x_c[n] \neq \pm\tau | y[n], \sigma_y^2, \sigma_w^2, \tau)} \geq 1 \end{cases} \tag{13}$$

In practice, the value of σ_w^2 is estimated from the observed speech by employing a VAD (*e.g.* [15]) and averaging the energy of the non-speech frames.



(a) With oracle knowledge of which samples are clipped.



(b) Blindly inferring which samples are clipped.

Figure 4: *Word error rates for automatic recognition of clipped speech ($\tau = P_{75}$) after declipping using the indicated algorithm. In the top panel, the knowledge of which samples are clipped is known* a priori. *In the bottom panel, the algorithm outlined in Sec. 3.3 is used to identify clipped samples before declipping.*

4. Declipping Techniques

In order to evaluate the practical effectiveness of the clipped sample estimation technique developed in Sec. 3.3, the method will be paired with two state-of-the-art declipping algorithms, which are briefly summarized here.

4.1. Consistent iterative hard thresholding

Kitic *et al.* recently proposed a highly-effective sparsity-based algorithm for declipping [9], which will be referred to as *Kitic-IHT*. Each incoming frame of clipped speech is represented using a sparse linear combination of Gabor basis vectors. The weights of the linear combination are learned using a modified form of Iterative Hard Thresholding [16]. The algorithm is deemed "consistent" as it requires the interpolated sample values to be greater than or equal to τ in the absolute sense, and carry the same sign as the corresponding clipped signal samples.

4.2. Constrained blind amplitude reconstruction

The *Constrained Blind Amplitude Reconstruction* (CBAR) algorithm [1] was recently proposed by the present authors. CBAR seeks to declip a signal by minimizing the energy of its second derivative, subject to the constraint that the interpolated samples agree with the sign of the underlying clipped signal and are greater than the clipping level, τ, in the absolute sense.

5. Results

5.1. Estimation accuracy

The accuracy of estimating τ in noise using the technique developed in Sec. 3.1 is depicted in Fig. 1. These data were generated by predicting τ over 500 independent trials of newly-generated white noise added to clipped speech at the indicated threshold. The markers in the plots show the mean value of the τ predictions over the trials. As can be seen, the technique is quite accurate for SNRs of $+10$ dB and above.

The accuracy of predicting the percentile value of τ, given the correct inferred amplitude value of τ for a particular speech waveform is shown in Fig. 2. Finally, the accuracy of predicting which samples are clipped, given perfect knowledge of the amplitude and percentile values of τ, using the technique presented in Sec. 3.3, is shown in Fig. 3. In agreement with intuition, the accuracy of clipped sample estimation decreases with both SNR and the clipping level.

5.2. Declipping

The real measure of the degree to which clipped sample estimation is potentially useful in practice is measured by how well a given declipping algorithm performs in noise when paired with the clipped sample estimation techniques.

The CMU Sphinx-III ASR system [17], trained on MFCC features [18] extracted from the clean (unclipped) DARPA RM1 database [19], was used to decode[2] speech that was clipped (at $\tau = P_{75}$) and added to noise at various intensities. Figure 4 depicts the corresponding WER as a function of SNR and using the indicated declipping algorithms, both when oracle (or perfect) knowledge of which samples were clipped is provided to the declining algorithm (upper panel) and when the methods developed in this paper are used to estimate which samples are clipped (lower panel). The benefit of the declipping algorithms becomes negligible at low SNRs when estimation of which samples are clipped is performed. Nevertheless, declipping remains quite effective at higher SNRs. At SNR = 15 dB, for example, the Kitic-IHT algorithm provides a 27.8% relative improvement over no declipping when paired with clipped sample estimation. At SNR = 20 dB, both CBAR and Kitic-IHT provide between 28% and 35% relative improvement over no declipping. While these improvements are (unsurprisingly) smaller than the improvements that could be obtained using oracle knowledge, they are nonetheless useful.

6. Conclusions

The results presented in this paper demonstrate the importance of robust methods to estimate the nature of the clipping processing when attempting to perform signal restoration from clipping in the presence of noise. As evidenced by the poor declipping results of state-of-the-art algorithms at low SNR compared to their performance in an oracle-knowledge situation (Fig. 4), declipping quality is only as good as the accuracy with which the clipped samples are estimated. This paper has introduced an effective technique for performing this estimation.

[2] Our configuration of Sphinx-III uses a standard bigram language model and 8-component GMM-based acoustic model. Sphinx-III is an HMM-based system. The MFCC features use a 40-band Mel-spaced triangular filter bank between 133 Hz and 6855 Hz. Windowing of 25.625 ms duration is performed at 100 frames per second using a Hamming window. Utterance-level cepstral mean subtraction is performed before training and decoding.

7. References

[1] M. Harvilla and R. Stern, "Least squares signal declipping for robust speech recognition," in *INTERSPEECH*, September 2014.

[2] A. Janssen, R. Veldhuis, and L. Vries, "Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes," *IEEE Trans. on Acoust., Speech and Signal Processing*, pp. 317–330, April 1986.

[3] J. Abel and J. Smith, "Restoring a clipped signal," in *IEEE Int. Conf. on Acoust., Speech and Signal Processing*, April 1991.

[4] W. Fong and S. Godsill, "Monte carlo smoothing for nonlinearly distorted signals," in *IEEE Int. Conf. on Acoust., Speech and Signal Processing*, May 2001.

[5] A. Dahimene, M. Noureddine, and A. Azrar, "A simple algorithm for the restoration of clipped speech signal," in *Informatica*, 2008, pp. 183–188.

[6] S. Miura, H. Nakajima, S. Miyabe, S. Makino, T. Yamada, and K. Nakadai, "Restoration of clipped audio signal using recursive vector projection," in *TENCON*, November 2011.

[7] A. Adler, V. Emiya, M. Jafari, M. Elad, R. Gribonval, and M. Plumbley, "Audio Inpainting," *IEEE Trans. on Acoust., Speech and Signal Processing*, pp. 922–932, April 2012.

[8] B. Defraene, N. Mansour, S. De Hertogh, T. van Waterschoot, M. Diehl, and M. Moonen, "Declipping of audio signals using perceptual compressed sensing," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 12, pp. 2627–2637, Dec 2013.

[9] S. Kitic, L. Jacques, N. Madhu, M. Hopwood, A. Spriet, and C. D. Vleeschouwer, "Consistent iterative hard thresholding for signal declipping," in *IEEE Int. Conf. on Acoust., Speech and Signal Processing*, May 2013.

[10] M. Akbacak, L. Burget, W. Wang, and J. v. Hout, "Rich system combination for keyword spotting in noisy and acoustically heterogeneous audio streams," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 8267–8271.

[11] P. Moreno, B. Raj, and R. Stern, "A vector taylor series approach for environment-independent speech recognition," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 2, May 1996, pp. 733–736 vol. 2.

[12] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 2, pp. 113–120, Apr 1979.

[13] A. Drake, *Fundamentals of Applied Probability Theory*. McGraw-Hill, Inc., 1967, ch. 1.

[14] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006, ch. 1.

[15] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *Signal Processing Letters, IEEE*, vol. 6, no. 1, pp. 1–3, Jan 1999.

[16] T. Blumensath and M. Davies, "Iterative thresholding for sparse approximations," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 629–654, 2008.

[17] C. S. S. Consortium, "CMU sphinx open source toolkit for speech recognition," http://cmusphinx.sourceforge.net/wiki/download/.

[18] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[19] P. Price, W. Fisher, J. Bernstein, and D. Pallett, "The DARPA 1000-word resource management database for continuous speech recognition," in *IEEE Int. Conf. on Acoust., Speech and Signal Processing*, April 1988.