MAXIMUM-LIKELIHOOD COMPENSATION OF ZERO-MEMORY NONLINEARITIES IN SPEECH SIGNALS

Robert W. Morris and Mark A. Clements

Center for Signal & Image Processing
Georgia Institute of Technology
Atlanta, Georgia 30332-0250
romorris@ece.gatech.edu, clements@ece.gatech.edu

ABSTRACT

In this paper, an algorithm to blindly compensate zero-memory nonlinear distortions of speech waveforms is derived and analyzed. This method finds a maximum-likelihood estimate of the distortion without *a priori* knowledge of the microphone characteristics by using the expectation-maximization algorithm. The autoregressive signal model coefficients are solved jointly with the nonlinearity estimate created by an extended Kalman filter. Also, a new family of nonlinear functions is developed for use with this algorithm, although the method can estimate the shape of any parametric zero-memory nonlinearity. These nonlinear distortions can degrade speech recognition rates, yet lower the perceptual quality only slightly. The compensation algorithm improves automatic speech recognition of distorted speech for a variety of such nonlinearities.

1. INTRODUCTION

It is well known that speech recognizers trained under one set of conditions do not perform well when used under other conditions. The training data is usually recorded in quiet environments with high-quality microphones. The new environment usually has a very complicated structure including noise, reverberation, and nonlinear effects. The nonlinear corruption can be created at several places including the amplifiers and microphones. In addition, the characteristics of this new environment may not be known prior to recognition.

Many methods have been developed to model and compensate for linear reverberation and additive noise. However, nonlinear effects on speech recognition applications have been less thoroughly investigated. Methods have included nonlinear compensation for speaker recognition applications [1] [2] and for general Gaussian input data [3]. In these studies, polynomial models trained on stereo training data [1] [3] and non-parametric, zero-memory models based on the cumulative density function of the speech [2] were tested. In this paper, a general method for estimating compressive and expansive nonlinearities using a parametric model with single-microphone data at recognition time is derived. By using a structure similar to iterative Wiener filtering [4], the algorithm produces a maximum-likelihood estimate of the nonlinearity parameters. A new family of functions for parameterizing nonlinearities is created for this algorithm. Also, the effect of compensating various nonlinearities on speech recognition accuracy is discussed.

This work was supported by the Texas Instruments Leadership University Program.

2. NONLINEARITY MODEL

To estimate the distortion on the speech, a model is necessary. The observation $y[n]$ is given by

$$y[n] = f_{\alpha,\beta}(s[n]), \tag{1}$$

where $s[n]$ is the input waveform, and $f_{\alpha,\beta}(x)$ is a memoryless nonlinearity parameterized by α and β. This nonlinearity must be constrained so that

$$\left.\frac{\partial}{\partial x} f_{\alpha,\beta}(x)\right|_{x=0} = 1 \;\; \forall \alpha, \beta. \tag{2}$$

It is also preferable for the derivatives of the function and its inverse to have a continuous, closed form. The function should also converge to a linear function for some α. The nonlinearity given by

$$f_{\alpha,\beta}(x) = \begin{cases} xe^{-\frac{1}{\beta}\mathrm{W}(\alpha\beta|x|^{\beta})} & \alpha > 0 \\ xe^{\alpha|x|^{\beta}} & \alpha \leq 0 \end{cases} \tag{3}$$

satisfies these constraints. The $\mathrm{W}(\cdot)$ term is Lambert's W function [5], which is the functional inverse of xe^{x} and can

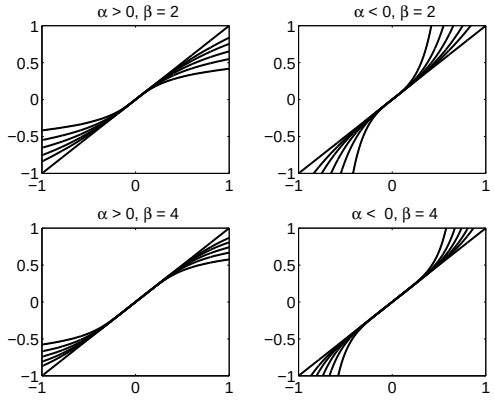

Figure 1: Examples of the nonlinearity

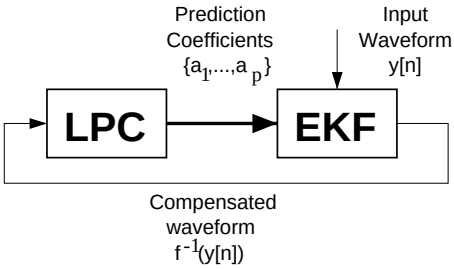

Figure 2: Block diagram of iterative algorithm

be approximated using the sum of a polynomial and a polynomial of $\log(1+x)$ given by

$$W(x) \approx \sum_{i=0}^{N} \alpha_i x^i + \sum_{j=1}^{M} \beta_j (\log(1+x))^j. \qquad (4)$$

This nonlinearity function has several advantages. First, its inverse can be written simply as $f^{-1}_{\alpha,\beta}(x) = f_{-\alpha,\beta}(x)$. In addition, the derivatives with respect to x, α, and β can all be written in closed form and are continuous across the domain, including α equal to zero. Several examples of possible functions are included in Figure 1. Within these plots, the magnitude of α ranges from 0.0 to 5.0, with the function becoming less linear as the magnitude of α increases.

3. ALGORITHM

The speech data is assumed to be autoregressive and stationary in blocks of 25 ms. Therefore, the signal can be modeled as

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + w[n]. \qquad (5)$$

By performing the nonlinearity on both sides of the equation and assuming that the residual remains a white noise process, the observations are expressed by

$$y[n] = f_{\mathbf{x}} \left(\sum_{k=1}^{p} a_k f^{-1}_{\mathbf{x}} (y[n-k]) \right) + v[n] \qquad (6)$$

where p is the vocal tract model order, a_k are the linear prediction coefficients, $f_{\mathbf{x}}(x)$ is the nonlinearity, and $f^{-1}_{\mathbf{x}}(x)$ is its inverse. The notation used for the nonlinearity is now parameterized by the vector $\mathbf{x}$. For the nonlinearity described above, $\mathbf{x}$ is equal to $(\alpha \ \beta)^T$. The goal of the algorithm is to jointly estimate the nonlinearity, $\mathbf{x}$, and the linear prediction coefficients, a_k, from the observations.

This joint estimation problem can be cast into the incomplete data form solved by the expectation-maximization (EM) algorithm. The observed data is $\mathbf{y}$ and the complete data, $\mathbf{z}$, is the actual speech waveform, $\mathbf{s}$, and the nonlinearity parameter, $\mathbf{x}$. The complete data has a pdf equal to $p(\mathbf{z}|\theta)$, where θ contains the model parameters that include the LP coefficients a_k, the residual variance σ^2, the nonlinearity mean $\hat{\mathbf{x}}$, and the nonlinearity covariance P_x. Since $\mathbf{s}$ and $\mathbf{x}$ are independent and share no model parameters, the processing of $p(\mathbf{x}|\theta)$ and $p(\mathbf{s}|\theta)$ can be done separately at each iteration. The sufficient statistics for describing the distribution $p(\mathbf{z}|\theta)$ are $r_{\mathbf{s}}$, the pth order autocorrelation sequence of $\mathbf{s}$, and the mean and covariance of $\mathbf{x}$.

During the expectation step, the sufficient statistics are estimated. The conditional autocorrelation is

$$r_{\mathbf{s}}(k) = E\left[\sum_{i=k}^{N} s[i]s[i-k] \middle| \mathbf{y}, \theta_i \right], \qquad (7)$$

where θ_i is the estimation of θ at the end of iteration i. By assuming that the variance of $\mathbf{x}$ is sufficiently small,

$$r_{\mathbf{s}}(k) \approx \sum_{i=k}^{N} f^{-1}_{\hat{\mathbf{x}}_i} (y[i])\, f^{-1}_{\hat{\mathbf{x}}_i} (y[i-k]). \qquad (8)$$

Next, the conditional expectation of the nonlinearity parameters,

$$\hat{\mathbf{x}} = E[\mathbf{x}|\mathbf{y}, \theta_i], \qquad (9)$$

is the minimum mean square estimator of $\mathbf{x}$. This can be approximated by using the extended Kalman filter (EKF). The nonlinear observation equation given by Equation 6 is linearized at each sample to derive the EKF [6].

In the maximization phase, the LP coefficients, $\hat{a}_{k,i+1}$ and σ^2_{i+1}, are calculated by using the autocorrelation method on the conditionally estimated $r_{\mathbf{s}}$.

The algorithm for estimating the nonlinearity for a single block is given in Table 1. In the single block algorithm, $\hat{\mathbf{x}}_i$ and $\hat{s}_i[n]$ are the estimates of the nonlinearity and the speech signal after iteration i. These are generated using $\hat{a}_{k,i}$ and σ^2_i, the estimated LPC coefficients at this iteration.

This algorithm has no dependence on the nonlinear model used, although all experiments in this paper use the nonlinearity described in Section 2. Since the nonlinearities have their largest effect on the high-energy segments, this algorithm is performed on the M highest energy blocks of the signal to be compensated.

Single block estimation and compensation algorithm:

Initialize

$$\hat{\mathbf{x}}_0(0) = \begin{pmatrix} \hat{\alpha}_0 \\ \hat{\beta}_0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \hat{s}_0[n] = y[n]$$

Iterate $i = 1, \ldots$

1) Autocorrelation method:

$$\sum_{l=1}^{p} \hat{a}_{k,i} r_{\hat{s}_i}(k-l) = -r_{\hat{s}_i}(k),\ k = 1, \ldots, p,$$

$$\sigma_i^2 = r_{\hat{s}_i}(0) + \sum_{l=1}^{p} \hat{a}_{k,i} r_{\hat{s}_i}(k)$$

2) Extended Kalman Filter, $n = 0, \ldots, N-1$:

$$\mathbf{K}(n) = \mathbf{P}(n)\mathbf{J}_f^T(n)\left(\mathbf{J}_f(n)\mathbf{P}(n)\mathbf{J}_f^T(n) + \sigma_i^2\right),$$

$$\nu(n) = y[n] - f_{\hat{\mathbf{x}}_i(n)}\left(\sum_{k=1}^{p} \hat{a}_{k,i} f^{-1}_{\hat{\mathbf{x}}_i(n)}(y[n-k])\right),$$

$$\hat{\mathbf{x}}_i(n+1) = \hat{\mathbf{x}}_i(n) + \mathbf{K}(n)\nu(n),$$

$$\mathbf{P}(n+1) = \left(\mathbf{I} - \mathbf{K}(n)\mathbf{J}_f(n)\right)\mathbf{P}(n),$$

$$\mathbf{J}_f(n) = \frac{\partial}{\partial \mathbf{x}}\left(f_{\mathbf{x}}\left(\sum_{k=1}^{p} \hat{a}_{k,i} f_{\mathbf{x}}^{-1}(y[n-k])\right)\right)\Bigg|_{\mathbf{x}=\hat{\mathbf{x}}_i(n)}$$

3) Estimate waveform:

$$\hat{s}_i[n] = f^{-1}_{\hat{\mathbf{x}}_i(N)}(y[n])$$

Until $\|\hat{\mathbf{x}}_i(N) - \hat{\mathbf{x}}_{i-1}(N)\|_2 < \delta$ **or** $i > L$

Table 1: Single block compensation algorithm

Let $\hat{s}^{(k)}[n]$ and $\hat{\mathbf{x}}^{(k)}$ be the final estimates of the signal and nonlinearity in block k, respectively. Because the all-pole model does not fit some phones such as nasals and nasalized vowels, the iteration does not always converge to reasonable values. These blocks are discarded to create a final working set.

To choose the parameters, a measure of the nonlinearity must be defined. Since the quantity of interest is the nonlinearity's effect on the waveform at the extreme magnitudes, this measure, ξ_k, is given by

$$\xi_k = \frac{\max|\hat{s}^{(k)}[n]|}{\max|y^{(k)}[n]|}. \tag{10}$$

Finally, the nonlinearity associated with the median value of this measurement is chosen

$$\hat{\mathbf{x}} = \hat{\mathbf{x}}^{\left(\arg\operatorname{median}_k \xi^k\right)}. \tag{11}$$

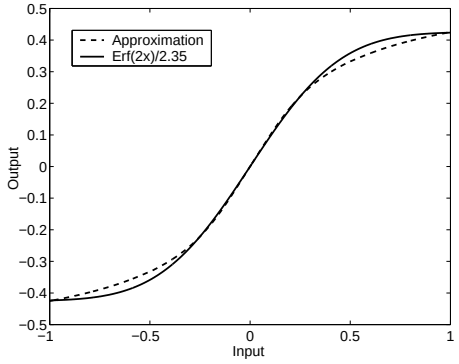

Figure 3: Estimation of Erf nonlinearity

The nonlinearity is assumed to be slowly varying, so this estimate can be used to compensate across the entire waveform to get

$$\hat{s}[n] = f_{\hat{\mathbf{x}}}^{-1}(y[n]). \tag{12}$$

Relevant recognition parameters, such as the mel-frequency cepstral coefficients, can be computed from this estimate.

4. EXPERIMENT

To provide a testing environment for different compensation algorithms, a connected digit recognizer was implemented using the automatic speech recognition software package created by the Institute for Signal and Information Processing at Mississippi State [7]. This system, which was trained using the TIDIGITS corpus, uses cross-word triphone models as the base unit. Each hidden Markov model consists of three states, while each continuous density state uses eight Gaussian mixtures. These models use mel-frequency cepstral coefficient (MFCC) feature vectors sampled every 10 ms. In addition, the front-end uses cepstral mean normalization (CMN) to remove any constant bias in the feature vectors. This system produces a word error rate (WER) of 0.9% on the TIDIGITS test set.

This method has been tested on speech corrupted by different nonlinearities. The function in Figure 3 is a nonlinearity based on the error function. The approximation was determined using the single block algorithm, showing the algorithm's ability to approximate nonlinear effects that do not fit the model exactly.

The algorithm was also tested for speech recognition by using different synthetic nonlinearities on the TIDIGITS data. For these experiments, the parameters M, L, p, and N were set to 20, 3, 8, and 200, respectively. The data was distorted using $f_{5,3}(x)$, an error function-based nonlinearity, and a hard limiter. These functions are shown in Figure 4. The signals generated by using $f_{-5,3}(x)$, the polynomial $0.2x + 0.4x^3 + 0.4x^5$, and a soft threshold tested the method

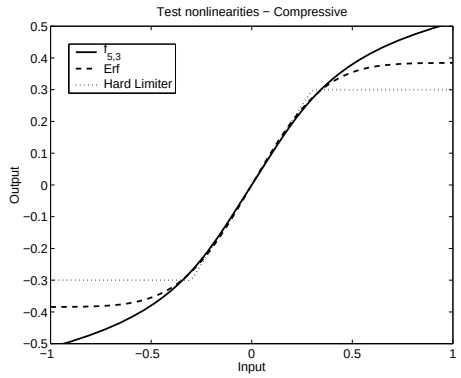

Figure 4: Compressive nonlinearities tested

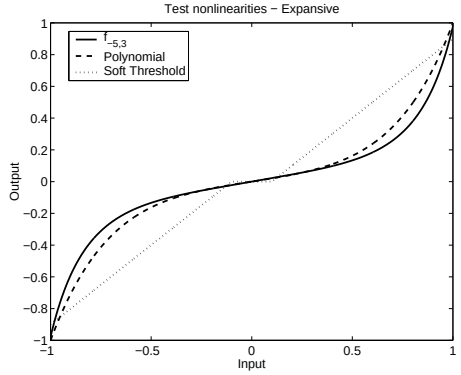

Figure 5: Expansive nonlinearities tested

on expansive nonlinearities. These functions are shown in Figure 5. The Itakura-Saito objective quality measurements for these different data sets are included in Table 3. Because the distortion levels are extremely varied across the waveform, both the average mean and maximum distances within the distorted corpus are calculated. The results of the recognition experiments are listed in Table 2. Since the distortion is nonlinear, there is no constant bias in the MFCCs. For this reason, the CMN algorithm improves the recognition rate only slightly. The compensation method effectively reduces the measured distortion and improves recognition rates in speech distorted by the smooth nonlinearities, but is far less effective at improving the limited and thresholded speech. This is expected, because no zero-memory inverse function exists to reconstruct these signals.

5. CONCLUSIONS

Zero-memory nonlinearities can have minor effects on the objective measure of speech quality, yet significantly degrade the error rates of automatic speech recognition systems. The presented algorithm enhances speech waveforms distorted by one-to-one nonlinearities and improves recognition rates without microphone-specific training data. Future work includes the investigation of additive and convolutional noise on this method, and the inclusion of these effects into the estimation framework.

Baseline Word Error Rate: 0.9%		
Nonlinearity	**Word Error Rate**	
	Original	**Compensated**
$f_{5,3}(x)$	2.7%	1.0%
$0.2\mathrm{Erf}(2.5x)$	2.9%	1.1%
Hard Limit	5.2%	4.6%
$f_{-5,3}(x)$	3.7%	2.1%
Polynomial	6.1%	2.2%
Soft Thresh	9.6%	9.8%

Table 2: Recognition using nonlinearity compensation

Nonlinearity	**Mean Distance**		**Max Distance**	
	Orig	**Comp**	**Orig**	**Comp**
$f_{5,3}(x)$	0.046	0.014	0.436	0.145
$0.2\mathrm{Erf}(2.5x)$	0.045	0.028	0.452	0.295
Hard Limit	0.087	0.103	0.838	0.965
$f_{-5,3}(x)$	0.129	0.059	1.10	0.473
Polynomial	0.172	0.053	1.31	0.254
Soft Thresh	0.199	0.207	2.05	2.05

Table 3: Itakura-Saito distance of speech

6. REFERENCES

[1] T. F. Quatieri, D. A. Reynolds, and G. C. O'Leary, "Estimation of handset nonlinearity with application to speaker recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 8, pp. 567–584, September 2000.

[2] R. Balchandran and R. J. Mammone, "Non-parametric estimation and correction of non-linear distortion in speech systems," in *Proc. ICASSP*, 1998, vol. 2, pp. 749–52.

[3] P. Celka, N. J. Bershad, and K. M. Vesin, "Analysis of stochastic gradient identification of polynomial nonlinear systems with memory," in *Proc. ICASSP*, 1999, vol. 3, pp. 1293–1296.

[4] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 26, pp. 197–210, June 1978.

[5] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert W function," *Advances in Computational Mathematics*, vol. 5, pp. 329–359, 1996.

[6] E. W. Kamen and J. K. Su, *Introduction to Optimal Esimation*, Springer, London, 1999.

[7] A. Ganapathiraju, N. Deshmukh, J. Hamaker, V. Mantha, Y. Wu, X. Zhang, J. Zhao, and J. Picone, "ISIP public domain LVCSR system," in *Proceedings of the Hub-5 Conversational Speech Recognition (LVCSR) Workshop*, June 1999.