NON-PARAMETRIC ESTIMATION AND CORRECTION OF NON-LINEAR DISTORTION IN SPEECH SYSTEMS

Rajesh Balchandran and Richard J. Mammone

CAIP Center, Rutgers University, New Jersey-08854

ABSTRACT

The performance of speech systems such as speaker recognition degrades drastically when there is mismatch between training and testing conditions, caused by non-linear distortion. This paper describes a technique to estimate and correct such non-linear distortion in speech. The focus is on constrained restoration of degraded speech, that is distortion in the test speech is undone relative to the training speech.

Restoration is a two step process - *estimation* followed by *inversion*. The non-linearity is estimated in the form of a look-up table by a process of statistical matching using a reference speech template. This statistical matching technique provides a very good estimate of the true non-linear characteristic, and the process is robust, computationally efficient, and universally applicable.

Speaker-ID experiments, using artificially corrupted test speech, showed significant improvement in performance after the test speech was 'cleaned' using this technique. The restoration process itself does not introduce appreciable distortion.

1. INTRODUCTION

Non-linear distortion is one of the prime causes of mismatch between training and testing conditions. Non-linearities are typically introduced in speech systems by carbon-button microphones, amplifiers, clipping and boosting circuits and automatic gain control circuits. Non-linearities distort the LP Cepstrum, the primary feature used for speaker recognition, resulting in poor performance. This paper presents a way to undo non-linearities in the time domain using a reference speech template.

2. NON-LINEARITY ESTIMATION BY STATISTICAL MATCHING [1]

The statistical matching technique [2] matches the distribution functions of clean and corrupted speech to estimate the non-linear process. This technique restricts the non-linear mapping to be, *zero memory, single valued* and *one-to-one* - that is, it expects a *point non-linearity*. A single valued, one-to-one, zero memory non-linearity can be expressed as:

$$Y[n] = g(X[n]) \tag{1}$$

where, X represents the clean speech samples which are the 'input' to the non-linear process $g(\cdot)$ and Y represents the corrupted speech samples at the 'output'. Let, $F_X(x)$ and $F_Y(y)$ be the cumulative distribution functions (CDF) of X and Y respectively and let $f_X(x)$, $f_Y(y)$ be their corresponding probability density functions.

Then, the input and output probability density functions (PDF) are related by the following theorem [2],

$$f_Y(y) = \frac{f_X(x)}{|dy/dx|} = \frac{f_X(x)}{|g'(x)|} \tag{2}$$

This theorem assumes that $(dy/dx) = g'(x)$ exists and is non-zero. The absolute value is required since the PDF is a non-negative function. The variable x on the right hand side should be replaced by its equivalent y that is, $x = g^{-1}(y)$ to obtain $f_Y(y)$ as a function of y.

2.1. Identification of Non-Linear System

The CDF of the input and output for any input x_0 and corresponding output $y_0 = g(x_0)$ are, by definition,

$$F_X(x_0) = P[X \leq x_0] \tag{3}$$

$$F_Y(y_0) = P[Y \leq y_0] \tag{4}$$

where, $P[\cdot]$ is the probability measure.

Therefore, a single valued, one-to-one and zero memory non-linearity can be identified by, *setting the CDF $F_X(x)$ equal to the CDF $F_Y(y)$*, that is, for any x_0 and its associated $y_0 = g(x_0)$,

$$F_Y(y_0) = \int_{-\infty}^{y_0} f_Y(y)dy = \int_{-\infty}^{x_0} f_X(x)dx = F_X(x_0) \tag{5}$$

Thus, Eq. (5) provides a technique to obtain the input-output characteristic of the non-linearity in the form of a look up table by mapping the CDFs of clean and corrupted speech.

2.2. CDF Matching Procedure

The CDF mapping process developed above can be carried out by the following two step procedure:

1. For each clean speech sample x_0, $F_X(x_0)$ at that point is determined from the CDF of the input.
2. The corresponding y_0 is found from the CDF of the output ($F_Y(y)$) such that,
$$F_Y(y_0) = F_X(x_0) \tag{6}$$

The set of all x_i and y_i obtained in this manner will be the lookup table that characterizes the non-linearity.

This research effort was made possible through the CAIP Center (Rutgers University) and by a grant from the U.S. Air Force at Rome Laboratories.

2.3. Implementation of The Mapping Process

The first requirement is to get an accurate estimate of the CDF from raw speech. The CDF of a random variable X is defined as,

$$F_X(x_i) \quad = \quad P[\,X \leq x_i\,] \tag{7}$$

The vector of data points can be easily converted into an unbiased estimator $Q_N(x)$ of the CDF of the probability distribution from which it was drawn [4]. If the N data points are located at values x_i, $(i = 1, 2, \ldots, N)$, then $Q_N(x_i)$ is the function giving the fraction of data points to the left of a given value x_i, that is,

$$Q_N(x_i) \quad \longleftrightarrow \quad P[\,X \leq x_i\,] \tag{8}$$

This function is constant between consecutive x_i's (when the x_i are sorted in ascending order) and jumps by the same constant $1/N$ at each x_i. Whenever there is a repeated data value, there is an additional jump. Thus, the sorted speech data forms the abscissa of the CDF and corresponding to the k^{th} sorted speech sample, the ordinate is given by,

$$F_X(x_k) \quad = \quad k/N \tag{9}$$

where, N is the number of speech samples.

The CDF of clean speech and the CDF of corrupted speech are obtained in the manner described above. The mapping process, can be represented graphically as shown in Figure 1.

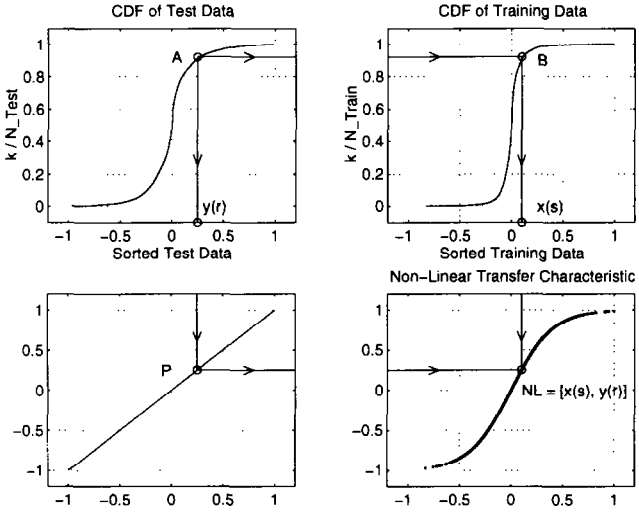

Figure 1: Non-Linearity Estimation by CDF Mapping

Essentially for every point on the CDF of test data (such as 'A') the corresponding point ('B') needs to be found. Mapping A onto B implies,

$$\begin{array}{rccc} & A & \longrightarrow & B \\ \Rightarrow & \frac{s}{N_{Train}} & \longleftrightarrow & \frac{r}{N_{Test}} \\ \Rightarrow & s & \longleftrightarrow & N_{Train}\,\frac{r}{N_{Test}} \end{array}$$

In practice, s may not correspond to an exact data point, so it is rounded to the nearest integer, that is,

$$s \quad = \quad (round)\left[\, N_{Train}\,\frac{r}{N_{Test}}\,\right] \tag{10}$$

And the point on the non-linearity curve is

$$NL \quad = \quad [\,x(s),\, y(r)\,] \tag{11}$$

2.4. Inverse Filtering of Corrupted Speech

The second stage in the non-linearity correction process is to invert the corrupted speech using the estimated transfer characteristic. The estimated non-linearity curve has, sorted test data on the Y axis and mapped training data on X axis. The inversion process can then be carried out by the following two steps:

1. For each test (corrupted) speech value, the closest match along the y-axis of the non-linearity curve is found.
2. Then, the corresponding value along the x-axis of the non-linearity curve will be the inverted speech value.

This inversion process is depicted graphically in Figure 2.

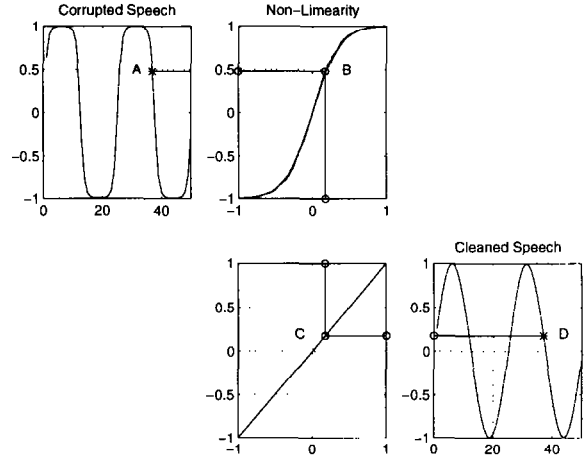

Figure 2: Speech Inversion Process

Since the points along the y-axis of the non-linearity curve are sorted test data values themselves, each value on the test (or corrupted) speech will have an 'exact' match on the y-axis of the non-linearity curve and therefore, the points can be found very efficiently using binary search. Thus, the CDF matching process facilitates very efficient and stable (since an exact match is guranteed) inversion.

3. EXPERIMENTS

The first set of experiments involved non-linearity estimation from a corrupted speech file and restoring the corrupted speech, while the second set consisted of speaker-ID experiments using 'clean' training data and non-linearity corrupted test data.

3.1. Generalized Training Data And Overestimation Compensation

The CDF matching technique would work the best when training (reference) and test data are from the same speaker, saying the same thing. But, this not practical, particularly for speaker ID, where the true identity of the speaker is not known and it is not possible to pick the correct training utterance.

To overcome this problem, it is necessary to have data that is 'general' enough to be used with any speaker or utterance. Speech typically has a Gamma distribution [3]. Therefore, the concatenation of a large number of utterances of sufficient length in the same amplitude range (say $[-1, 1]$) and having the same mean should also be Gamma distributed. For our experiments, generalized training data was obtained using sentences from the 'TIMIT' database. Since the data is to be sorted, the concatenated utterance can be down-sampled after sorting to get rid of redundant information, and reduce computational cost. The training data so obtained, is general enough to be used as the reference signal for non-linearity estimation.

Clean speech was artificially corrupted using the *sigmoid* non-linearity and subjected to the estimation process. The sigmoid non-linearity [1] is given by,

$$y = \frac{1}{1 + e^{-\mu x}} \tag{12}$$

As the parameter μ is increased the non-linearity approaches a hard-limiter. Figure 3 shows the estimated (marked 'Estimation without Compensation') and actual non-linearity ($\mu = 5$) curves.

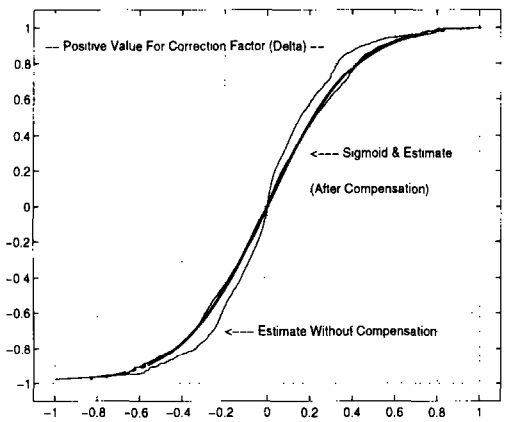

Figure 3: Sigmoid ($\mu = 5$) Estimation After Overestimation Compensation

Clearly the 'estimation without Compensation' curve is not as close to the actual curve as desired. This deviation or overestimation is primarily due to the 'general' nature of the reference speech, which is not from the same speaker.

The overestimation problem can be corrected by 'shifting' the estimated curve towards the 45° line. This can be carried out during the initial estimation step itself - needing no additional computation. The estimate for the non-linearity point $x(s)$ given in Eq. 11 is modified as follows:

$$x_{shifted}(s) = x(s) - \delta\,[y(s) - x(s)] \tag{13}$$

where, δ is a correction factor that controls the extent of compensation. The difference term $(y(s) - x(s))$ measures the distance of the estimated point from the $y = x$ or 45^0 line and the amount or extent of compensation is determined by its magnitude. The correction factor δ provides an additional degree of control which can be used to fine tune the amount of compensation. Typically, δ ranges between 0 and 1. A large value of δ is used when there is minimal non-linearity, that is, when the curve is nearly linear and small value of δ is used when the curve is highly non-linear. For the sigmoid with $\mu = 5$, $\delta = 0.3$ was found to be the most effective.

As can be seen in Figure 3 the estimate after application of the compensation process is seen to nearly overlap the true non-linearity curve.

To see if the estimation process itself introduces any distortion, non-linearity estimation is carried out on clean speech. This 'linear' case is shown in Figure 4.

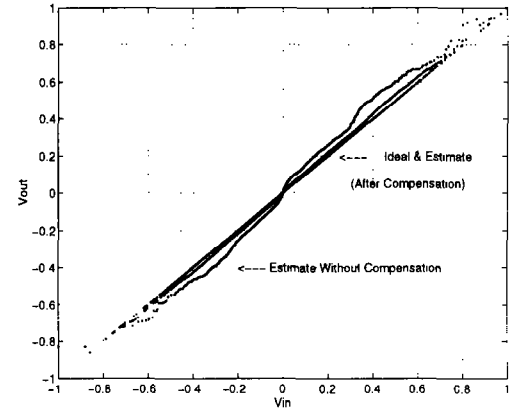

Figure 4: 'Non-Linearity' Estimation For Linear Case

As can be seen from the figure the estimate after compensation is very close to the 45° line showing that the statistical matching process itself does not introduce appreciable distortion.

Figures 5 and 6 show the estimated non-linearity and restored speech for the the sigmoid ($\mu = 5$) and the cubic ($y = x^3$).

The RMS error $\left[\sqrt{\frac{1}{N}\sum_{n=0}^{N-1}(s(n) - \hat{s}(n))^2}\right]$ between clean and corrupted and between clean and restored speech for these two cases is given in Table 1.

The figures and the RMS error values show that the restoration is very effective.

3.2. Speaker-ID Experiments

Speaker-ID experiments were carried out using data from the 'TIMIT' database. Speakers under the 'train' section, of the database from the "New England" dialect were used for all the experiments. In all, 38 speakers (24 male and 14 female) were considered with 5 training and 5 test utterances for each. The LP derived cepstrum of order 12 was used as

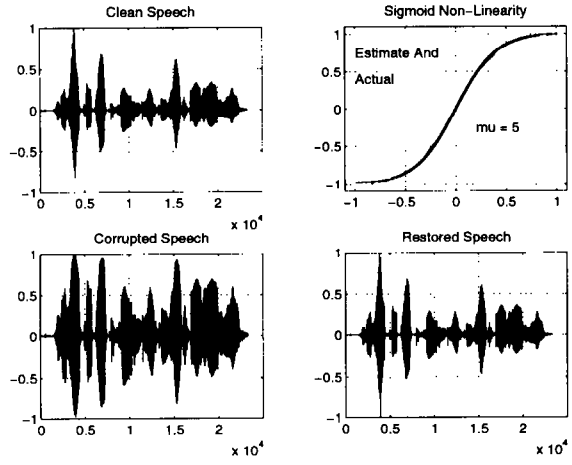

Figure 5: Non-Linearity Estimation And Speech Restoration - Sigmoid ($\mu = 5$)

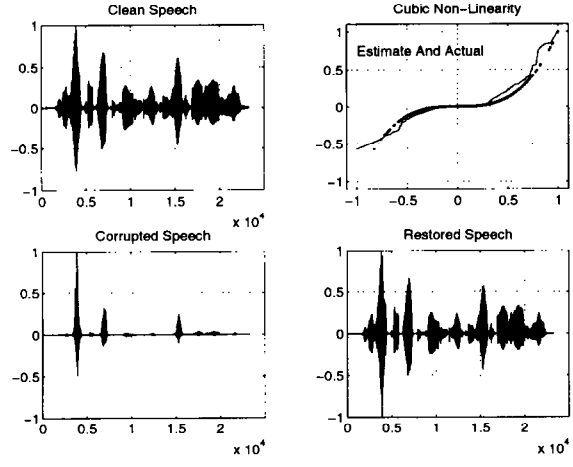

Figure 6: Non-Linearity Estimation And Speech Restoration - Cubic

the feature vector and Vector Quantization (Codebook size 64) was used for classification [1].

The base system performance was 95.26 %. The first set of experiments consisted of training the system with clean speech and testing with speech corrupted using the sigmoid non-linearity for $\mu = 2, 5, 10$ & 20. For the next set, the same trained codebooks were used, but the corrupted speech was 'cleaned' using the statistical matching technique before testing. The results are shown in Table 2.

From the results we see that there is substantial improvement in speaker-ID performance after restoration by the statistical matching technique. The improvement is about 0.5% for $\mu = 2$ and is more pronounced at about 51% for $\mu = 20$. Complete restoration does not occur for very high values of μ like 20. This is because at large μ values, the non-linearity is almost like a hard-limiter making restoration very difficult.

Non-Linearity	RMS Error (%)	
	Before Restoration	After Restoration
Sigmoid ($\mu = 5$)	12.0	0.70
Cubic	10.0	2.0

Table 1: RMS Error Measurements : [Clean Distorted], [Clean - Restored] Speech.

Base-System Performance : 95.26 %		
Sigmoid Parameter	Speaker ID Performance	
	Before Rstn.	After Rstn.
$\mu = 2$	94.7 %	95.2 %
$\mu = 5$	74.7 %	91.1 %
$\mu = 10$	71.6 %	88.4 %
$\mu = 20$	53.2 %	80.5 %

Table 2: Speaker ID Performance Before And After Restoration

4. CONCLUSIONS

From the experiments, it can be concluded that the statistical matching technique is a very practical, efficient and effective way to estimate and compensate point non-linearities in speech systems. The process itself does not introduce appreciable distortion.

In situations where training data is corrupted the restoration process can be carried out before training also. In this case both training and test data will be restored relative to another clean data set. The process need not be restricted to the sigmoid or cubic non-linearities – experiments [1] have shown that it can be used to correct any non-linearity or even a combination of non-linearities.

5. REFERENCES

[1] Rajesh Balchandran. Non-parametric estimation and correction of non-linear distortion in speech systems. Master's thesis, Rutgers University, May, 1997.

[2] Julius S. Bendat. *Nonlinear System Analysis And Identification From Random Data.* John Wiley & Sons, 1990.

[3] L. R. Rabiner and Ronald W. Schafer. *Digital Processing of Speech Signals.* Prentice-Hall, Inc., Englewood Cliffs, New Jersey 07632, 1978.

[4] Saul A. Teukolsky et al William H. Press. *Numerical Recipies in C: The art of scientific computing.* Cambridge University Press, 1994.