USE OF A RELIABILITY COEFFICIENT IN NOISE CANCELLING BY NEURAL NET AND WEIGHTED MATCHING ALGORITHMS

Nestor Becerra Yoma, Fergus McInnes, Mervyn Jack*
Centre for Communication Interface Research, University of Edinburgh
80 South Bridge, Edinburgh EH1 1HN, U.K.
E-Mail:nestor@ccir.ed.ac.uk

ABSTRACT

The problems of efficacy estimation in noise cancelling by a neural net (LIN-Lateral Inhibition Net [5]) and the use of this information in weighting matching algorithms are focused. Since the effect of noise on the speech signal is variable and the backpropagation training algorithm is essentially stochastic (most common patterns have more influence in the weights re-estimation process), it is reasonable to suppose that the LIN efficacy depends on the input and each noisy frame could be associated to a *reliability* coefficient that attempts to measure how reliable is the result of the neural net processing. Isolated word recognition experiments have shown that *reliability* weighting can result in a mean error rate reduction as high as 96, 80, 58 and 36 % at SNR=12, 6, 3 and 0dB, respectively, when the noise is white Gaussian.

1. INTRODUCTION

Many techniques have been proposed to try to solve the noise sensitivity of speech recognition algorithms. Many of these noise cancellation methods can be seen as a system that processes a noisy input and produces an output with the influence of noise reduced. In all the approaches seen so far, the efficacy of noise cancelling systems has been considered constant in the sense that it is supposed that the noisy speech signals are processed equally, independently of the segmental SNR and the power spectral distribution. In [5], a technique based on a noise cancelling neural net (LIN) and segmental SNR weighting in acousting pattern matching algorithm was proposed. However, further experiments showed that the improvement of SNR weighting depended on LIN training conditions and suggested that the neural net noise cancellation efficacy should be included in the weighting procedure.

The contributions of this paper concern: a) the estimation of *reliability* in noise cancelling by a neural net; and b) combination of this reliability estimation with weighted matching algorithms. This approach has not been found in the literature and seems sufficiently generic to be employed with other noise reduction techniques. A modified backpropagation algorithm, that can reduce the number of iterations needed to train LIN, is also proposed.

*Supported by a grant from CNPq-Brasilia/Brasil

2. LATERAL INHIBITION NET (LIN): A NOISE CANCELLATION NEURAL NET

Masking is basically the suppression of the lowest by the highest spectral components. Lateral inhibition is one of the processes responsible for the masking phenomena in different sensory systems and this concept was used to train the noise reduction neural network, LIN (Lateral Inhibition Net), employed in this research. Given:

- E_j, the normalised log energy at the output of the filter j in a bank of N filters;
- $F_i^c = (E_1^c, E_2^c, E_3^c, ..., E_N^c)$, frame i of clean signal;
- $F_i^n = (E_1^n, E_2^n, E_3^n, ..., E_N^n)$, frame i after it has noise added;

the lateral inhibition function (LI) can be set as:

$$LI(E_j) \quad = \quad E_j + f(E_1, E_2, E_3, ..., E_N) \qquad (1)$$

where the function $LI()$ was approximated with multilayer perceptrons with one hidden layer. The output function for the hidden layer nodes was $\sigma(x) = 1/(1+e^{-x})$ and the output function for input and output layers is linear. Each input node receives the energy of one filter and the same energy is fedforward to the output node in order to compound the equation (1). The number of input, hidden and output nodes were equal to the number of filters N [5]. The LIN was trained with the following conditions that define the lateral inhibition function:

$$LI(F_i^n) \approx LI(F_i^c)$$

$$LI(F_i^c) \approx F_i^c$$

All the weights of the neural net (except those on the feedforward connections from the inputs to the outputs which were always equal to 1) were estimated with the classical backpropagation algorithm with cross-validation [1]. The training data were made up of input-reference pattern pairs. Initially, the reference patterns were frames of clean signal, F_i^c, and the input patterns were generated adding white Gaussian noise to F_i^c at 4 different SNR's (Clean, 18dB, 12dB, and 6dB). Therefore, each frame F_i^c originated 4 training input-reference pairs. In a modified version of the training algorithm, $LI(F_i^c)$ was used instead of F_i^c as reference patterns.

2.1. LIN Training Database

Sounds that present low energy (typically fricatives) are the first to be masked by corrupting signals, and using these speech frames as training patterns could mean learning the neural network with an information that is lost even for moderate SNRs. In the results reported in this paper, energy was used as discriminative parameter. Initially the maximum energy of the utterance was computed and then all the frames that were below a given threshold from the maximum energy were discarded. According to some preliminary experiments a suitable threshold would be 25dB.

3. LIN AND RELIABILITY IN NOISE REDUCTION

In recognition tests, reference (clean utterances) and testing patterns (noisy utterances) are processed by LIN, and hence in the acoustic pattern matching algorithm the local distances correspond to $d[LI(F_k^c), LI(F_i^n)]$ instead of $d[F_k^c, F_i^n]$, where k denotes a reference frame and i a test one. In the experiments reported here, the distance function d was the Euclidean metric.

A noise cancelling neural net can be seen as a system that processes a noisy input and produces an output with the influence of noise reduced. Since there are several levels of distortions and the backpropagation training algorithm is essentially stochastic (most common patterns have more influence in the weights re-estimation process), it is reasonable to suppose that the LIN efficacy depends on the input and each noisy frame could be associated to a *reliability* coefficient that attempts to measure how reliable is the result of LIN processing. Due to the fact that the noise cancelling depends on $d[LI(F_i^c), LI(F_i^n)]$ (the smaller this distance is, the better is the noise influence cancelling), the *reliability* coefficient (r) could be related to this distortion by means of the following function:

$$r = \begin{cases} 1 & \text{if } d[LI(F_i^c), LI(F_i^n)] \leq \delta \\ \frac{\delta}{d[LI(F_i^c), LI(F_i^n)]} & \text{if } d[LI(F_i^c), LI(F_i^n)] > \delta \end{cases}$$

At the recognition procedure, the clean version F_i^c of the noisy testing frame F_i^n is not available but, due to the fact that the power spectral distribution of the corrupting signal is known (white Gaussian noise), F_i^c can be set as a function of F_i^n and the local SNR. After LIN has been trained, the training data-base could be used to approximate the relation between $d[LI(F_i^c), LI(F_i^n)]$, and F_i^n and the local SNR. Consequently, if the segmental SNR could be computed frame by frame and given that F_i^n is available, the *reliability* coefficient could be estimated frame by frame during the recognition process.

3.1. Local SNR Estimation

If the noise is poorly correlated and uncorrelated with the speech signal, it is possible to estimate the power of the clean speech from the autocorrelation function of the noisy signal [5]. Given that $R_x(m)$, $R_s(m)$ and $R_n(m)$ are the autocorrelation functions of the noisy speech, the clean speech and the noise signals, respectively, the following coeffficient can be computed frame by frame:

$$n = \frac{R_s(0)}{R_x(0)} = \frac{R_s(0)}{R_n(0) + R_s(0)} \quad (2)$$

where $n = 1$ if $segmentalSNR = \infty$; and $n = 0$ if $segmentalSNR = -\infty$. $R_s(0)$ is estimated by means of applying some properties of the autocorrelation function and quadratic interpolation [5]:

$$R_s(0) = \frac{4 \times R_x(1) - R_x(2)}{3} \quad (3)$$

The coefficient n can be computed frame by frame because it needs just the autocorrelation of the noisy signal at points m=0, 1 and 2. Observe that the estimation of the noise power in silence intervals is not needed and the method captures the dynamics of the speech and noise signals' energy. The segmental SNR and the coefficient n are related by the following equation:

$$n = \frac{10^{SNR/10}}{1 + 10^{SNR/10}} \quad (4)$$

3.2. Mean Distortions

As an approximation, it can be assumed that the distortion $d[LI(F_i^c), LI(F_i^n)]$ depends exclusively on the local SNR. The mean-distortion for each SNR may be estimated at the LIN training procedure. Then, during the recognition process, the distortion $d[LI(F_i^c), LI(F_i^n)]$ for the frame F_i^n with local SNR equal to snr can be approximated by the mean-distortion $\overline{D_{snr}}$ at local SNR equal to snr. $\overline{D_{snr}}$ can be computed for some SNR's at the LIN training procedure and, by means of linear interpolation, it can be estimated for other values of SNR.

For the results presented in this paper, $\overline{D_{snr}}$ was computed for SNR=18, 12, 6, 3 and 0dB by employing the LIN evaluation database, after LIN had been trained. During the recognition procedure, the coefficient n was estimated by means of the autocorrelation function (2)(3) and the curve

$\overline{D_{snr}} \times localSNR$ was mapped into the n domain by using the equation (4). The constant δ was made equal to 0.004, a value that was shown to be suitable according to some tests.

4. MODIFIED BACKPROPAGATION ALGORITHM

In the ordinary neural net training algorithm, the quadratic error is computed between the reference F_i^c and the output $LI(F_i^n)$. However, the efficacy of LIN is related to the distortion $d[LI(F_i^c), LI(F_i^n)]$: the smaller $d[LI(F_i^c), LI(F_i^n)]$ is, the smaller should be the recognition error rate. As a consequence, it can be interesting to include the condition of minimization of $d[LI(F_i^c), LI(F_i^n)]$ in the training algorithm in a more explicit way. The minimisation of $d[F_i^c, LI(F_i^n)]$ leads to the reduction of $d[LI(F_i^c), LI(F_i^n)]$, but this distance also depends on the angle between $LI(F_i^c) - F_i^c$ and $LI(F_i^n) - F_i^c$. In the modified algorithm, the clean signal F_i^c was replaced with $LI(F_i^c)$ as the reference for the noisy frames, and the quadratic error was computed between the reference $LI(F_i^c)$ and the output $LI(F_i^n)$.

At the ordinary LIN training algorithm (BLT, Backpropagation LIN Training), in each epoch the backpropagation minimizes the quadratic error of the following sequence of pairs reference-output: a) F_i^c and $LI(F_i^c)$; and b) F_i^c and $LI(F_i^n)$, for all the local SNR's included in the training database.

At the modified training algorithm (MLT, Modified LIN Training), in each epoch the backpropagation minimizes the quadratic error of the following sequence of pairs reference-output: a) F_i^c and $LI(F_i^c)$; and b) $LI(F_i^c)$ and $LI(F_i^n)$, for all the local SNR's included in the training database. This is more coherent with the conditions that define the lateral inhibition function (see section 2). It is interesting to highlight that the reference is not constant as in the ordinary backpropagation algorithm, but it is modified iteration by iteration because $LI(F_i^c)$ depends on LIN, and LIN's weights are re-estimated each time that a reference-output pair is presented to the training algorithm.

5. EXPERIMENTS OF WORD RECOGNITION

5.1. Database and Pre-processing

The proposed methods were tested with speaker-dependent isolated word (English digits from 0 to 9) recognition experiments. The tests were carried out employing the two speakers (one female and one male) from the Noisex database. The isolated clean words were automatically end detected and generated the database used in this research. For each speaker, the 100 training clean utterances (10 repetitions per digit) generated 10 reference sets (set of repetition 1 of each word, set of repetition 2 of each word, etc). The 100 testing clean utterances were used to create the noisy database, as explained with more details in [5], by adding white noise at 6 global-SNR levels: clean speech, +18dB, +12dB, +6dB, +3dB and 0dB. Before the Gaussian noise was added, the speech signals were low pass filtered, using a 10th order Tchebychev filter with cut off frequency equal to 3700 Hz, and down sampled from 16000 to 8000 samples/sec. The band from 300 to 3400 Hz was covered with 14 Mel 2nd order IIR digital filters. The normalised log energy of each filter was an input of LIN as explained in section 2. After LIN processing 10 cepstral coefficients were computed.

5.2. Training the Neural Net

For each speaker, the frames from the set of repetition 1 of the training database (section 5.1) generated the input-reference pattern pairs used by the LIN training algorithm to estimate the weights. The frames from the set of repetition 2 of the training database generated the input-reference pattern pairs used for evaluation of the performance of LIN. Several training conditions (learning rate, initial weights and data base) were tested and the one that gave the best results on the test data was chosen. For each speaker, the LIN training variables were kept constant in order to compare the MLT and BLT algorithms at the same conditions.

5.3. Results

The results presented in this paper were achieved with 1000 recognition tests for each SNR: 10 reference sets $\times$ 100 testing utterances. The following configurations were tested: the ordinary DTW algorithm [2] with cepstral coefficient without (DP-C) and with (DP-L) LIN processing; the proposed weighted DP algorithm [5] with LIN processing, (DPW-$\overline{D}$) with the mean-distortions method for *reliability* estimation and (DPW-n) with local SNR weighting; and finally, the two-step DP matching [3] with LIN, (DP2-$\overline{D}$) with the mean-distortions method for *reliability* estimation and (DP2-n) with local SNR weighting. The recognition error rates are presented in tables 1 and 2 for the female speaker, and in tables 3 and 4 for the male one. For the female speaker, LIN was trained with 6132 and 3869 iterations with the BLT and MLT algorithms, respectively. For the male speaker, LIN was trained with 7403 (BLT) and 1702 (MLT) iterations.

6. DISCUSSION

LIN showed a substantial reduction in error rates even without reliability weighting. LIN with the ordinary DTW algorithm ($DP - L$) practically eliminated the influence of the noise at SNR=18 and 12 dB, and resulted in a mean reduction of 86, 69 and 48% at SNR=6, 3 and 0dB, respectively. Moreover, the error introduced for clean testing signals was almost zero.

As can be seen in tables 1 and 2 (female speaker) and tables 3 and 4 (male speaker), the *reliability* coefficient estimated with the mean-distortion method gave a greater reduction in the error rate than the SNR weighting in all noisy conditions. When LIN was trained by means of the MLT algorithm, the reduction due to *reliability* weighting was as high as 100, 83

and 56% at SNR=12, 6 and 3dB, respectively, while the SNR weighting resulted in a much smaller reduction in most of the cases and even in an increase of the error rate in other cases.

Table 1: Recognition error rate (%) for the female speaker. LIN was trained with the BLT algorithm.

SNR	*Clean*	*18dB*	*12dB*	*6dB*	*3dB*	*0dB*
DP-C	0.1	3.5	31.9	67.0	70.6	75.6
DP-L	0.1	0.1	1.2	11.5	31.9	53.5
DPW-$\overline{D}$	0.2	0.1	0.1	4.0	17.2	33.3
DPW-n	0.1	0.4	1.0	6.4	26.0	43.9
DP2-$\overline{D}$	0.1	0.0	0.1	3.5	15.9	32.1
DP2-n	0.1	0.2	0.4	4.0	20.6	38.2

Table 2: Recognition error rate (%) for the female speaker. LIN was trained with the MLT algorithms.

SNR	*Clean*	*18dB*	*12dB*	*6dB*	*3dB*	*0dB*
DP-C	0.1	3.5	31.9	67.0	70.6	75.6
DP-L	0.1	0.2	0.6	5.9	10.6	24.5
DPW-$\overline{D}$	0.1	0.0	0.0	0.7	6.1	17.9
DPW-n	0.1	0.1	0.3	3.0	9.5	30.1
DP2-$\overline{D}$	0.1	0.0	0.0	0.5	5.6	17.6
DP2-n	0.1	0.1	0.1	2.3	6.5	23.6

Table 3: Recognition error rate (%) for the male speaker. LIN was trained with the BLT algorithm.

SNR	*Clean*	*18dB*	*12dB*	*6dB*	*3dB*	*0dB*
DP-C	0.0	16.8	49,9	65.1	69.4	74.6
DP-L	0.3	0.4	1.6	9.8	21.9	41.8
DPW-$\overline{D}$	0.1	0.1	0.1	1.3	6.3	24.6
DPW-n	0.5	0.5	3.4	10.4	20.6	43.4
DP2-$\overline{D}$	0.1	0.1	0.2	1.2	6.6	25.2
DP2-n	0.3	0.1	1.7	8.2	17.2	38.9

Table 4: Recognition error rate (%) for the male speaker. LIN was trained with the MLT algorithm.

SNR	*Clean*	*18dB*	*12dB*	*6dB*	*3dB*	*0dB*
DP-C	0.0	16.8	49,9	65.1	69.4	74.6
DP-L	0.0	0.6	2.7	9.2	22.6	38.2
DPW-$\overline{D}$	0.1	0.0	0.0	2.2	7.8	24.1
DPW-n	0.5	0.8	3.3	11.5	22.0	36.1
DP2-$\overline{D}$	0.1	0.0	0.0	2.3	7.8	24.8
DP2-n	0.3	0.0	0.9	8.5	17.7	31.6

The one-step weighted algorithm proposed in [5] showed almost the same performance as the two-step one [3] with the *reliability* coefficient, but resulted in a poorer improvement when the SNR estimation was used as a weighting parameter. This must be due to the fact that in the one-step algorithm the influence of a frame on decisions must be proportional to its weighting coefficient, and the *reliability* coefficient includes not only the information concerning the segmental SNR, but also the LIN characteristic in the form of the mean-distortion curve, and provides a more accurate estimation of the reliability of the information extracted from each frame.

According to tables 1-4, the *reliability* coefficient as a weighting parameter gave the best results, with the MLT algorithm for the female speaker and with the BLT algorithm for the male one. However, the MLT algorithm kept the error rate below 2.5% at SNR=6dB and below 10% at SNR=3dB for both speakers.

7. CONCLUSIONS

The combination of LIN and weighted DP algorithms proved to be effective in reducing the influence of white Gaussian noise, and the error introduced for clean testing signals was almost zero. The *reliability* coefficient gave better results than the SNR estimation as a weighting parameter and this must arise from the fact that this coefficient takes into account not only the local SNR estimation but also the characteristic response of LIN in the form of the mean-distortion curve. The weighted DP algorithms helped to reduce the error rate, but its improvement decreased when the SNR became more severe. The one-step DP matching proposed in [5] was also shown to be effective in reducing the error rate, and led to approximately the same error rates as the two-step matching [3] when the *reliability* weighting was used. The *reliability* coefficient as a weighting parameter seems to be a generic approach and could be employed with other noise cancelling techniques.

A drawback of LIN is the strong influence of training conditions (learning rate, initial weights and data-base) in the final results and several configurations had to be tested. In this sense, the inclusion of the *reliability* coefficient seems to be an important advance because it caused a reduction of the error rate in all the cases, independently of the training configurations. Future work includes the generalization of LIN structure to other types of noises, adaptation to new environments and a more generic estimation for the *reliability* coefficient.

References

[1] S.Haykin. *Neural Networks, A Comprehensive Foundation* Macmillan College Publishing Company. 1994.

[2] X.D.Huang, Y.Ariki, M.A.Jack. *Hidden Markov Models for Speech Recognition.* Edinburgh University Press, 1990.

[3] Hidefumi Kobatake, Yousuke Matsunoo. *Degraded Word Recognition Based on Segmental Signal-to-Noise Ratio Weighting.* ICASSP 1994, Vol. I, pp.425-428.

[4] A.Varga, H.J.M.Steeneken, M.Tomlinson and D.Jones. *The Noisex-92 study on the effect of additive noise in automatic speech recognition.* Technical report, DRA Speech Research Unit, U.K., 1992.

[5] N.B.Yoma, F.R.McInnes, M.A.Jack. *Improved Algorithms for Speech Recognition in Noise Using Lateral Inhibition and SNR Weighting.* Eurospeech'95, pp.461-464.