WEIGHTED MATCHING ALGORITHMS AND RELIABILITY IN NOISE CANCELLING BY SPECTRAL SUBTRACTION

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

ABSTRACT

This paper addresses the problem of speech recognition with signals corrupted by additive noise at moderate SNR. A technique based on spectral subtraction and noise cancellation reliability weighting in acoustic pattern matching algorithms is studied. A model for additive noise is proposed and used to compute the variance of the hidden clean signal information and the reliability of the spectral subtraction process. The results presented in this paper show that a proper weight on the information provided by static parameters can substantially reduce the error rate.

1. INTRODUCTION

Due to the fact that the intervals with highest energies are less corrupted by additive noise, it is reasonable to suppose that these intervals provide more reliable information for speech recognition than those intervals with lower energies. In [1] and [2] were proposed two weighted matching algorithms to take into account the local SNR. Both algorithms were tested with poorly correlated and white Gaussian noises but with different noise cancellation techniques. In [2] these two algorithms were tested in combination with a noise cancellation neural net and it was shown they could reduce the error rate. However, further experiments showed that the improvements due to the weighted Dynamic Programming algorithms depended on the neural net training conditions, and suggested that the weighting coefficient $w(t)$ should take into account not only the segmental SNR but the characteristics of the noise reduction method. Following this idea, in [3] was proposed the use of a weighting parameter based on reliability in noise cancelling. This parameter takes into account not only the local SNR but also the characteristic response of the noise cancellation method in the form of a mean distortion curve [3].

The contributions of this paper concern: a)combination of weighted matching algorithms with spectral subtraction (SS) technique; and b) analysis of SS in terms of reliability in noise cancelling. The approach covered by this paper has not been found in the literature and seems to be generic and interesting from the practical applications point of view. In this exploratory research, the techniques were tested with DTW recognition algorithms on an isolated word recognition task. DTW was used because it is a simple and generic algorithm which allows many noise cancelling techniques to be compared without the need for extensive tuning of the modelling. However, the authors believe the tecniques explored here could also be employed by a weighted Viterbi (HMM) algorithm previously proposed in [2].

*Supported by a grant from CNPq-Brasilia/Brasil

2. SINUSOIDAL MODEL FOR ADDITIVE NOISE

Given that $s(i)$, $n(i)$ and $x(i)$ are the clean speech, the noise and the resulting noisy signal, respectively, the additiveness condition may be set as:

$$x(i) = s(i) + n(i) \quad (1)$$

In the results presented in this paper, the signal was processed by 14 Mel filters. At the output of filter j the noisy signal is given by:

$$x_j(i) = s_j(i) + n_j(i) \quad (2)$$

and its mean energy in a frame by:

$$\overline{x_j^2(i)} = \overline{s_j^2(i)} + \overline{n_j^2(i)} + \overline{2s_j(i)n_j(i)} \quad (3)$$

where $\overline{x_j^2(i)} = \frac{1}{N}\sum_{i=1}^{N} x_j^2(i)$, $\overline{s_j^2(i)} = \frac{1}{N}\sum_{i=1}^{N} s_j^2(i)$, $\overline{n_j^2(i)} = \frac{1}{N}\sum_{i=1}^{N} n_j^2(i)$, $\overline{2s_j(i)n_j(i)} = \frac{1}{N}\sum_{i=1}^{N} 2s_j(i)n_j(i)$ and N is the length of the frames in number of samples.

If the speech signal and the noise are uncorrelated, $E(\overline{2s_j(i)n_j(i)}) = 0$ in a long term analysis, where $E()$ corresponds to the expected value. However, the condition $\overline{2s_j(i)n_j(i)} = 0$ may not be satisfied in a short term analysis (i.e. a 25 ms frame) and the noise is certainly not perfectly stationary. Consequently, once the noise is added the clean signal energy, $\overline{s_j^2(i)}$, becomes a hidden information and cannot be recovered with a 100% accuracy. As a result, $\overline{s_j^2(i)}$ should be treated as a stochastic variable and could be associated to a variance that indicates how accurate is the estimation of the clean signal energy.

Initially, the signals $s_j(i)$ and $n_j(i)$ are considered sinusoidal components with frequency f_j, the central frequency of filter j, with a phase difference ϕ. Under these assumptions,

$$\overline{x_j^2(i)} = \frac{a_{s_j}^2}{2} + \overline{n_j^2(i)} + a_{s_j}a_{n_j}cos(\phi) \quad (4)$$

where a_{s_j} and a_{n_j} are the amplitudes of the speech signal and noise components respectively: $\overline{s_j^2(i)} = a_{s_j}^2/2$ and $\overline{n_j^2(i)} = a_{n_j}^2/2$.

3. CORRECTION OF THE SINUSOIDAL MODEL

The sinusoidal model for additive noise represented by equation (4) assumes that the components $s_j(i)$ and $n_j(i)$ at the output of filter j have frequency f_j and a phase difference ϕ in a given frame. These assumptions are not perfectly accurate in practice. Firstly, the 14 mel filters are not highly selective, which reduces the validity of the assumption of coherence between both components. Secondly, the phase ϕ between $s_j(i)$ and $n_j(i)$ is not necessarily constant and a few discontinuities in the phase difference may occur, although many of them are unlikely in a short term analysis (i.e. a 25 ms frame). However, the sinusoidal model represents the fact that there is a variance in the short term analysis and specifies the relation between this variance and the clean and noise signal levels. Due to the lack of coherence between $s_j(i)$ and $n_j(i)$ and to the discontinuity in the phase difference, the variance predicted by the model is higher than the the real one for the same frame length, and a correction should be included in (4). According to (4) and considering that the random variable ϕ was uniformly distributed between $-\pi$ and π:

$$Var[\overline{x_j^2(i)}|\overline{s_j^2(i)},\overline{n_j^2(i)}] = 0.5a_{s_j}^2 a_{n_j}^2$$

In order to estimate the correction of the sinusoidal model, the coefficient r_j defined as

$$r_j = \frac{\overline{2s_j(i)n_j(i)}}{a_{s_j}a_{n_j}} \quad (5)$$

was computed with clean speech and only-noise frames. According to (5), $Var[r_j|\overline{s_j^2(i)},\overline{n_j^2(i)}]$ should be equal to 0.5 but due to the lack of coherence between $s_j(i)$ and $n_j(i)$ and to the discontinuity in the phase difference,

$$Var[r_j|\overline{s_j^2(i)},\overline{n_j^2(i)}] < 0.5$$

and a correction factor k_j needs to be included in (4):

$$\overline{x_j^2(i)} = \frac{a_{s_j}^2}{2} + \overline{n_j^2(i)} + a_{s_j}a_{n_j}\sqrt{k_j}cos(\phi) \quad (6)$$

where k_j is defined as

$$k_j = 2Var[r_j|\overline{s_j^2(i)},\overline{n_j^2(i)}]$$

4. CHANNEL VARIANCE

With the sinusoidal model for additive noise represented by (6), the variance (or uncertainty) of the hidden information $\overline{s_j^2(i)}$ given the observed information $\overline{x_j^2(i)}$ is estimated. Solving (6) for a_{s_j} and using $\overline{s_j^2(i)} = a_{s_j}^2/2$:

$$\overline{s_j^2(i)} = a_{n_j}^2 k_j cos^2(\phi) + \overline{x_j^2(i)} - \overline{n_j^2(i)} - a_{n_j}\sqrt{k_j}cos(\phi)\sqrt{a_{n_j}^2 k_j cos^2(\phi) + 2(\overline{x_j^2(i)} - \overline{n_j^2(i)})} \quad (7)$$

The equation above sets $\overline{s_j^2(i)}$ as a function of ϕ, $\overline{n_j^2(i)}$ and $\overline{x_j^2(i)}$:

$$\overline{s_j^2(i)} = g(\phi, \overline{n_j^2(i)}, \overline{x_j^2(i)}) \quad (8)$$

The function $g(\phi, \overline{n_j^2(i)}, \overline{x_j^2(i)})$ was used to estimate $Var[\log(\overline{s_j^2(i)})|\overline{x_j^2(i)}]$ considering that the random variable ϕ was uniformly distributed between $-\pi$ and π and that $\overline{n_j^2(i)}$ is concentrated near its mean $E[\overline{n_j^2(i)}]$. The variance $Var[\log(\overline{s_j^2(i)})|\overline{x_j^2(i)}]$ is given by:

$$Var[\log(\overline{s_j^2(i)})|\overline{x_j^2(i)}] = E[\log^2(\overline{s_j^2(i)})|\overline{x_j^2(i)}] - E^2[\log(\overline{s_j^2(i)})|\overline{x_j^2(i)}]$$

where

$$E[\log^2(\overline{s_j^2(i)})|\overline{x_j^2(i)}] \simeq \frac{1}{2\pi}\int_{-\pi}^{\pi}\log^2[g(\phi, E[\overline{n_j^2(i)}], \overline{x_j^2(i)})]d\phi$$

and

$$E[\log(\overline{s_j^2(i)})|\overline{x_j^2(i)}] \simeq \frac{1}{2\pi}\int_{-\pi}^{\pi}\log[g(\phi, E[\overline{n_j^2(i)}], \overline{x_j^2(i)})]d\phi$$

The integrals for estimating $E[\log^2(\overline{s_j^2(i)})|\overline{x_j^2(i)}]$ and $E[\log(\overline{s_j^2(i)})|\overline{x_j^2(i)}]$ were computed by means of Simpson's rule with the interval $(-\pi, \pi)$ divided in 100 regular partitions. The difference $\overline{x_j^2(i)} - \overline{n_j^2(i)}$ in (7) was replaced with $\sigma(Est(\overline{s_j^2(i)}))$ (see section 5) when evaluating g().

5. SPECTRAL SUBTRACTION

Spectral subtraction (SS) may be defined as

$$Est(\overline{s_j^2(i)}) = \overline{x_j^2(i)} - E(\overline{n_j^2(i)}) \quad (9)$$

where $Est(\overline{s_j^2(i)})$ is the estimation of the clean signal energy and $E(\overline{n_j^2(i)})$ is the mean noise energy estimation made in non-speech intervals. Due to the fact that $\overline{2s_j(i)n_j(i)} = 0$ may not be true in a short term analysis and that the noise energy presents fluctuations, $Est(\overline{s_j^2(i)})$ may be negative in those channels with low SNR. In order to avoid negative magnitude estimates a rectifying function $\sigma()$is applied:

$$\sigma(Est(\overline{s_j^2(i)})) = \begin{cases} Est(\overline{s_j^2(i)}) & \text{if } Est(\overline{s_j^2(i)}) \geq \epsilon \\ \epsilon & \text{if } Est(\overline{s_j^2(i)}) < \epsilon \end{cases} \quad (10)$$

where ϵ is an arbitrary low constant.

6. WEIGHTED MATCHING ALGORITHMS

Some modifications were included in matching algorithms in order to weight the reliability of the information extracted from testing frames. A weighting coefficient $w(t)$ ($w(t) = 1$, maximum reliability; $w(t) = 0$, minimum reliability) is associated to each testing frame in order to be employed in the modified vesions of the DTW and Viterbi (HMM) algorithms [2]. The main idea behind the modifications made on Viterbi (HMM) and DTW algorithms is that the influence of a frame on decisions must be proportional to its coefficient $w(t)$. The proposed one-step weighted DP algorithm was compared with the two-step DP algorithm proposed in [1].

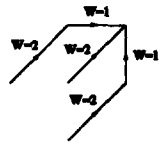

Figure 1. Local condition.

The proposed one-step DP equation that corresponds to the local condition shown in Fig.1 is given as follows :

$$G(t,r) = min \left(\begin{array}{c} \frac{G(t-2,r-1)W(t-2,r-1)+2w(t-1)d(t-1,r)+w(t)d(t,r)}{W(t-2,r-1)+2w(t-1)+w(t)} \\ \frac{G(t-1,r-1)W(t-1,r-1)+2w(t)d(t,r)}{W(t-1,r-1)+2w(t)} \\ \frac{G(t-1,r-2)W(t-1,r-2)+2w(t)d(t,r-1)+w(t)d(t,r)}{W(t-1,r-2)+3w(t)} \end{array} \right)$$

and

$$W(t,r) = \left\{ \begin{array}{c} W(t-2,r-1)+2w(t-1)+w(t) \\ W(t-1,r-1)+2w(t) \\ W(t-1,r-2)+3w(t) \end{array} \right.$$

This DP equation takes into account the weight $w(t)$ frame by frame, and the calculation of the overall distance, $G(t,r)$, is affected by $d(t,r)$ according to $w(t)$: if $w(t) = 1$ (high *reliability* or local SNR), the weight of $d(t,r)$ is maximum; if $w(t) = 0$ (very low *reliability* or local SNR), the importance of $d(t,r)$ is zero.

The algorithm proposed in [1] consists of the following two-step processing. Firstly, the optimal alignment path $c_k = (t_k, r_k), k = 1, 2, ..., K$ is obtained using the ordinary DP matching algorithm, where t_k and r_k are the frame numbers of the testing and reference patterns respectively. The second step is the calculation of the global distance between the utterances weighted by $w(t_k)$ along the optimal path obtained at the first step.

7. RELIABILITY IN NOISE CANCELLING

It is reasonable to suppose that the uncertainty related to SS in a channel would be proportional to $Var[\log(\overline{s_j^2(i)})|\overline{x_j^2(i)}]$: the higher $Var[\log(\overline{s_j^2(i)})|\overline{x_j^2(i)}]$ is, the less reliable is the information provided by $Est(\overline{s_j^2(i)})$; and the lower this variance is, the higher is the probability of $Est(\overline{s_j^2(i)})$ being close to the clean signal information $\overline{s_j^2(i)}$. The weighting coefficient $w(t)$ [2] [3], to be used by the weighted algorithms (section 6) and that attempts to measure how reliable is the result of the noise cancelling method in a frame, could be related to the mean $Var[\log(\overline{s_j^2(i)})|\overline{x_j^2(i)}]$ in all the channels by means of the following function (Fig. 2) [3]:

$$w(t) = \left\{ \begin{array}{ll} 1 & \text{if } MeanVar \leq \delta \\ \frac{\delta}{MeanVar} & \text{if } MeanVar > \delta \end{array} \right. \quad (11)$$

where

$$MeanVar = \frac{1}{14} \sum_{j=1}^{14} Var[\log(\overline{s_j^2(i)})|\overline{x_j^2(i)}] \quad (12)$$

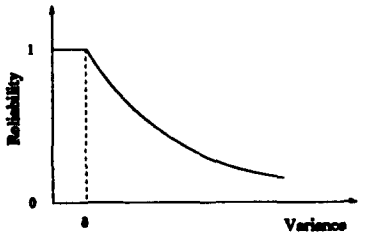

Figure 2. Reliability coefficient vs variance.

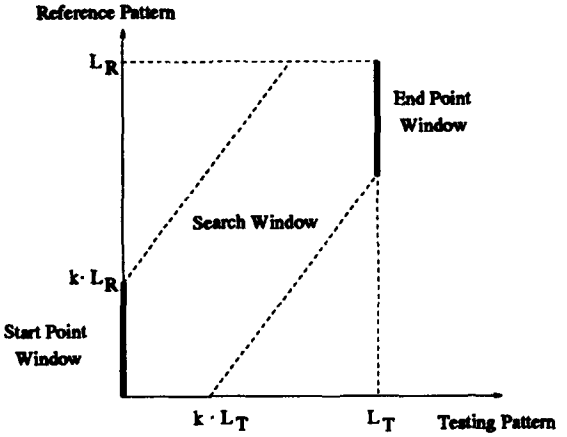

Figure 3. End-point constraints relaxation.

8. END POINT RELAXATION

The reliability in noise cancelling weighting was tested by means of isolated word Dynamic Time Warping algorithms. The isolated words were automatically end detected using an algorithm based on autoregressive analysis of noise [4] and the average length of the testing utterances decreases as the SNR gets more severe. Consequently, the endpoint constraints on the DP algorithms were relaxed by means of opening up the ends of the search region allowing the alignment path to start by comparing the first frame of the testing pattern with any of the first reference frames inside the search window, and to end by comparing the last test frame with any of the last reference frames inside the search window (see Fig. 3). Due to the fact that the length of the testing utterances presented a high variation, the sides of the search window were made proportional to the utterance length.

9. EXPERIMENTS

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 [5].

The signals were low pass filtered by using a filter with cut off frequency 3700 Hz, down sampled from 16000 to 8000 samples/sec, and high-pass filtered by employing a filter with cut off frequency 120 Hz. The data signal was divided in 25ms frames with 12.5ms overlapping. Each frame was processed with a Hamming window before the spectral estimation. The band from 300 to 3400 Hz was covered with 14 Mel 2nd order IIR digital filters. At the output of each channel the energy was computed and SS was applied. Finally, 10 cepstral coefficients were computed.

The results presented in this paper were achieved with

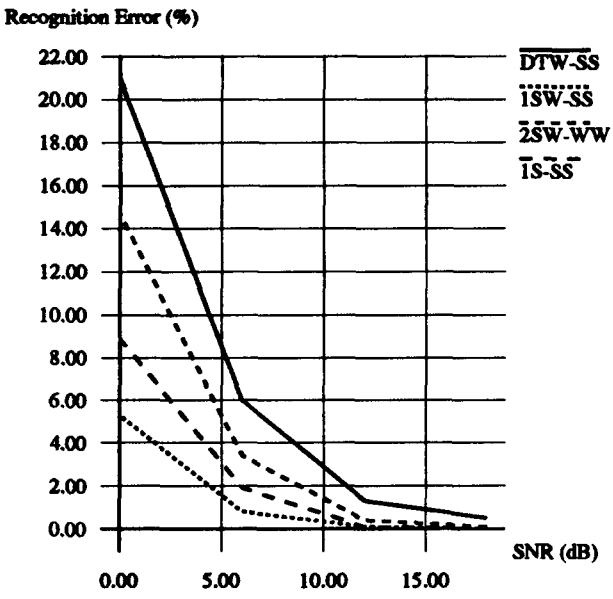

Figure 4. Results for the car noise (Noisex database).

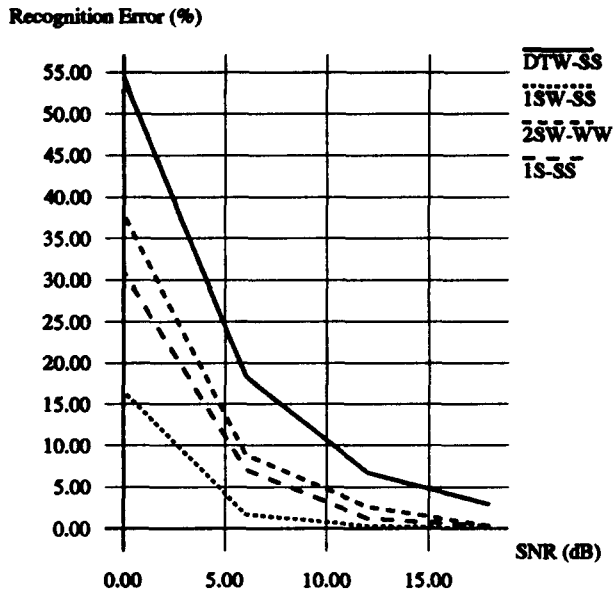

Figure 5. Results for the speech noise (Noisex database).

1000 recognition tests for each SNR. The following configurations were tested: the ordinary DTW algorithm [6] with SS (DTW-SS); the proposed one-step weighted DP algorithm [2] with SS (1SW-SS); the two-step DP matching [1] (2SW-SS) also with SS; and finally, the proposed one-step DP algorithm with SS but without reliability in noise cancelling weighting, $w(t) = 1$ (1S-SS). The constant δ was made equal to 43, a value that was shown to be suitable according to some tests. For each configuration several search window widths, k (Fig. 3), were tested and the one that gave minimum error rate was chosen to plot the graphs shown in Figs. 4 and 5.

10. DISCUSSION AND CONCLUSION

As can be seen in Figs. 4 and 5, the one step algorithm in combination with the noise cancellation reliability weighting gave the lowest error rate. This reduction in the error rate was due to a) the ability of the one step algorithm in normalising the overall distance to the length of the alignment path, and b) the information provided by the noise cancellation reliability coefficient. The ordinary DTW does not take into consideration which point of the start window the optimal alignment path begins and was very sensitive to the search window width. Consequently, when k was increased (Fig. 3), DTW-SS and 2SW-SS increased the error rate after reaching an optimum search window. On the other hand, the DP equation shown in section 6 computes the overall weight $W(i,j)$ step-by-step and was almost independent to the alignment path length. As a result, 1SW-SS should be compared with 1S-SS in order to separate the improvement due to the alignment path normalisation and the one due to the noise cancelling reliability weighting.

When compared with 1S-SS, 1SW-SS showed reductions of 58% and 40% in the error rate at SNR=6dB and 0dB for the car noise. At SNR=18dB and 12dB both configurations gave error rate equal to 0 and 0.1%, repectively. For the speech noise, 1SW-SS presented reductions of 75%, 76% and 46% at SNR=12, 6 and 0dB. At SNR=18dB the error rate went from 0.3% to 0. As can be seen, the improvement due to the reliability weighting was higher for the speech noise than for the car one. This must result from the facts that the speech noise is less stationary than the car noise so the estimation of noise energy is less accurate, and that the reliability coefficient is also a function of the local SNR so low energy intervals have low weight in the pattern matching process. Therefore, noise cancellation reliability weighting made the SS process more robust to variations in the noise stationarity.

REFERENCES

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

[2] 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.

[3] N.B.Yoma, F.R.McInnes, M.A.Jack. *Use of a Reliability coefficient in noise cancelling by Neural Net and Weighted Matching Algorithms.* ICSLP'96, pp. 2297-2300.

[4] N.B.Yoma, F.R.McInnes, M.A.Jack. *Robust speech pulse detection using adaptive noise modelling.* IEE Electronics Letters, Vol. 32, No. 15, July 1996, pp. 1350-1352.

[5] 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.

[6] H. Sakoe, S. Chiba. *Dynamic Programming Algorithm Optimization for Spoken Word Recognition.* IEEE Trans. on ASSP, vol. ASSP-26, No 1, Feb. 1978, pp. 43-49