SPECTRAL SUBTRACTION AND MEAN NORMALIZATION IN THE CONTEXT OF WEIGHTED MATCHING ALGORITHMS

Nestor Becerra Yoma[*], *Fergus R. McInnes, Mervyn A. 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

Additive and convolutional noises are the main problems to be solved in order to make speech recognition successful in real applications. A model for additive noise is used to deduce a spectral subtraction (SS) estimation and to show that the channel transfer function could be effectively removed after the additive noise being cancelled by SS. Then, SS and mean normalization are tested in combination with a weighting procedure to reduce the influence of the rectifying function. All the experiments were done in the context of weighted matching algorithms and the approaches proved effective in cancelling both additive noise and the transmission channel function.

1. INTRODUCTION

In [1], a model for additive noise using IIR filters was proposed and used to compute the reliability related to the spectral subtraction (SS) process. The reliability in noise cancelling was used to weight DP algorithms and shown to be useful in reducing the error rate. Nevertheless, the low selectivity of the IIR filters made the system more vulnerable to convolutional distortions and the use of a DFT bank filter is desirable because it provides an infinite rejection outside the filter band.

If there is only convolutional distortion, a widely used technique is Cepstral Mean Normalizatoin (CMN). CMN is effective and efficient but its behaviour is hard to predict when additive noise is also present [2].

The contributions of this paper concern: a)the generalisation of the model for additive noise for the case of DFT filters; b)the proof that under some conditions, the log of the SS estimation is equal to the expected value of the log of the clean signal energy; and c) the proof that the effect of an unmatched transmission channel can effectively be removed by means of the classic mean normalization technique after SS. 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. The results presented in this paper provide a theoretical justification for the use of mean normalization after SS, and show that both techniques in combination with weighted matching algorithms can effectively remove both additive and convolutional distortions.

[*]Supported by a grant from CNPq-Brasilia/Brasil

2. MODEL FOR ADDITIVE NOISE USING DFT FILTERS

Given that $s(i)$, $n(i)$ and $x(i)$ are the clean speech, the noise and the resulting noisy signal, respectively, the additiveness condition in the temporal domain 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 DFT Mel filters. If $S(k)$, $N(k)$ and $X(k)$ correspond to the FFT transform of $s(i)$, $n(i)$ and $x(i)$ at the point k, and ϕ_k is the phase difference between $S(k)$ and $N(k)$, the additiveness condition is then set by:

$$X(k) = S(k) + N(k) \quad (2)$$

According to the cosine rule,

$$|X(k)|^2 = |S(k)|^2 + |N(k)|^2 + 2 \cdot |S(k)| \cdot |N(k)| \cdot cos(\phi_k) \quad (3)$$

The energy at the output of the filter m, $\overline{x_m^2}$, is computed by means of:

$$\overline{x_m^2} = \sum_{k \in filter\ m} G(m,k) \cdot |X(k)|^2 \quad (4)$$

where $G(m,k)$ is the set of weights that define the filter m. If $|X(k)|^2$ in (4) is replaced with the expression given in (3), $\overline{x_m^2}$ can be set as

$$\overline{x_m^2} = \overline{s_m^2} + \overline{n_m^2} + \sum_{k \in filter\ m} 2 \cdot G(m,k) \cdot |S(k)| \cdot |N(k)| \cdot cos(\phi_k) \quad (5)$$

where: $\overline{s_m^2}$ and $\overline{n_m^2}$ are the filter m mean frame energy of the clean speech and noise signal, respectively.

Assuming that the phase difference $\phi(k) = \phi$, $N(k)$ and $X(k)$ are considered constant inside each one of the 14 DFT Mel filters indexed by m:

$$\overline{x_m^2} = \overline{s_m^2} + \overline{n_m^2} + 2 \cdot \sqrt{\overline{s_m^2}} \cdot \sqrt{\overline{n_m^2}} \cdot cos(\phi) \quad (6)$$

The model for additive noise represented by (6) assumes that the components $|S(k)|$ and $|N(k)|$ and the phase difference ϕ are constant inside every filter in a given frame. These assumptions are not perfectly accurate in practice. Firstly, the 14 DFT mel filters are not highly selective, which reduces the validity of the assumption of low variation of these parameters inside the filters. Secondly, the phase ϕ between $|S(k)|$ and $|N(k)|$ 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, this 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 these approximations the variance predicted by the model is higher than the the real one for the same frame length, and a correction should be included. Using (6) and considering that ϕ was uniformly distributed between $-\pi$ and π:

$$Var[\frac{\overline{x_m^2}}{2} | \overline{s_m^2}, \overline{n_m^2}] = 0.5 \cdot \overline{s_m^2} \cdot \overline{n_m^2}$$

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

$$k_m = \frac{\overline{x_m^2} - \overline{s_m^2} - \overline{n_m^2}}{2 \cdot \sqrt{\overline{s_m^2}} \sqrt{\overline{n_m^2}}} \quad (7)$$

was computed with clean speech and only-noise frames. According to (6), $Var[k_m | \overline{s_m^2}, \overline{n_m^2}]$ should be equal to 0.5 but due to the approximations this variance is lower than 0.5 and a correction factor c_m needs to be included in (6):

$$\overline{x_m^2} = \overline{s_m^2} + \overline{n_m^2} + 2 \cdot \sqrt{c_m} \cdot \sqrt{\overline{s_m^2}} \cdot \sqrt{\overline{n_m^2}} \cdot cos(\phi) \quad (8)$$

where c_m is defined as

$$c_m = 2Var[k_m | \overline{s_m^2}, \overline{n_m^2}]$$

Solving (8) for $\overline{s_m^2}$

$$\overline{s_m^2}(\phi, \overline{n_m^2}, \overline{x_m^2}) = 2 \cdot A^2 \cdot cos^2(\phi) + B - 2 \cdot A \cdot cos(\phi) \cdot \sqrt{A^2 \cdot cos^2(\phi) + B} \quad (9)$$

where $A = \sqrt{\overline{n_m^2} c_m}$ and $B = \overline{x_m^2} - \overline{n_m^2}$.

3. CHANNEL VARIANCE AND RELIABILITY IN NOISE CANCELLING BY SS

With the model for additive noise represented by (9), the variance (or uncertainty) of the hidden information $\overline{s_m^2}$ given the observed information $\overline{x_m^2}$ is estimated in the logarithmic domain assuming that the random variables ϕ and $\overline{n_m^2}$ are uncorrelated, ϕ is uniformly distributed between $-\pi$ and π and that $\overline{n_m^2}$ is concentrated near its mean $E[\overline{n_m^2}]$. The variance $Var[\log(\overline{s_m^2}) | \overline{x_m^2}]$ is given by:

$$Var[\log(\overline{s_m^2}) | \overline{x_m^2}] = E[\log^2(\overline{s_m^2}) | \overline{x_m^2}] - E^2[\log(\overline{s_m^2}) | \overline{x_m^2}] \quad (10)$$

where

$$E[\log^2(\overline{s_m^2}) | \overline{x_m^2}] \simeq \frac{1}{2\pi} \int_{-\pi}^{\pi} \log^2[\overline{s_m^2}(\phi, E[\overline{n_m^2}], \overline{x_m^2})] d\phi$$

$$E[\log(\overline{s_m^2}) | \overline{x_j^2}] \simeq \frac{1}{2\pi} \int_{-\pi}^{\pi} \log[\overline{s_m^2}(\phi, E[\overline{n_m^2}], \overline{x_m^2})] d\phi$$

Equation (16) below suggests that the expected value of the hidden information $\log(\overline{s_m^2})$ is approximately equal to the log of the spectral subtraction (SS) estimation ($Est(\overline{s_m^2})$) if $Est(\overline{s_m^2}) = \overline{x_m^2} - E[(\overline{n_m^2})]$, where $E[\overline{n_m^2}]$ is the mean noise energy estimation made in non-speech intervals. In order to avoid negative magnitude estimates a rectifying function $r()$ is applied:

$$r(Est(\overline{s_m^2}), \varepsilon) = \begin{cases} Est(\overline{s_m^2}) & \text{if } Est(\overline{s_m^2}) \geq \varepsilon \\ \varepsilon & \text{if } Est(\overline{s_m^2}) < \varepsilon \end{cases} \quad (11)$$

where ε is an arbitrary low constant. As in [1] the weighting coefficient w, to be used by the weighted algorithms [1] and that attempts to measure how reliable is the result of the noise cancelling method in a frame, was defined as:

$$w = \begin{cases} 1 & \text{if } TotalVar \leq \delta \\ \frac{\delta}{TotalVar} & \text{if } TotalVar > \delta \end{cases} \quad (12)$$

where

$$TotalVar = \sum_{m=1}^{14} Var[\log(\overline{s_m^2}) | \overline{x_m^2}] \quad (13)$$

$Var[\log(\overline{s_m^2}) | \overline{x_m^2}]$ was estimated by means of $E[\log(\overline{s_m^2}) | \overline{x_j^2}] \simeq \log[\overline{x_m^2} - E[\overline{n_m^2}]]$ (see section 4) and the integral for estimating $E[\log^2(\overline{s_m^2}) | \overline{x_m^2}]$ was computed by means of Simpson's rule with the interval $(-\pi, \pi)$ divided in 100 regular partitions and replacing the difference $B = \overline{x_m^2} - \overline{n_m^2}$ in (9) with $r(Est(\overline{s_m^2}), \varepsilon)$.

4. ADDITIVE AND CONVOLUTIONAL NOISE CANCELLING

Given the model for additive noise represented by (9), the expected value of $\log[\overline{s_m^2}(\phi, \overline{n_m^2}, \overline{x_m^2})]$ given the observed information $\overline{x_m^2}$ would be:

$$E[\log(\overline{s_m^2}) | \overline{x_m^2}] = E[\log(2 \cdot \frac{A^2}{B} \cdot cos^2(\phi) + 1 - 2 \cdot \frac{A}{B} \cdot cos(\phi) \cdot \sqrt{A^2 \cdot cos^2(\phi) + B}) | \overline{x_m^2}] + E[\log(B) | \overline{x_m^2}] \quad (14)$$

Assuming again that the random variables ϕ and $E[\overline{n_m^2}]$ are uncorrelated, ϕ is uniformly distributed between $-\pi$ and π and that $\overline{n_m^2}$ is concentrated near its mean $E[\overline{n_m^2}]$, $E[\log(\overline{s_m^2})|\overline{x_m^2}]$ can be written as:

$$E[\log(\overline{s_m^2})|\overline{x_j^2}] \simeq \frac{1}{\pi}\int_{-\pi}^{\pi}\log[\sqrt{\frac{(\bar{A})^2\cdot cos^2(\phi)}{\bar{B}}+1}-\frac{\bar{A}}{\sqrt{\bar{B}}}\cdot cos(\phi)]d\phi+\log(\bar{B}) \quad (15)$$

where $\bar{A}=\sqrt{E[\overline{n_m^2}]\cdot c_m}$ and $\bar{B}=\overline{x_m^2}-E[\overline{n_m^2}]$. Replacing the variable ϕ with $u=-\frac{\bar{A}}{\sqrt{\bar{B}}}\cdot cos(\phi)$, the integral in (2) becomes

$$\frac{1}{\pi}\int_{-\pi}^{\pi}\log[\sqrt{\frac{(\bar{A})^2\cdot cos^2(\phi)}{\bar{B}}+1}-\frac{\bar{A}}{\sqrt{\bar{B}}}\cdot cos(\phi)]d\phi = \frac{2\cdot\sqrt{\bar{B}}}{\bar{A}\cdot ln(10)\cdot\pi}\int_{-\frac{\bar{A}}{\sqrt{\bar{B}}}}^{\frac{\bar{A}}{\sqrt{\bar{B}}}}\frac{sinh^{-1}(u)}{\sqrt{1-\frac{\bar{B}}{(\bar{A})^2}\cdot u^2}}du=0$$

because the functions $sinh^{-1}(u)$ and $\sqrt{1-\frac{\bar{B}}{(\bar{A})^2}\cdot u^2}$ are odd and even respectively. Consequently,

$$E[\log(\overline{s_m^2})|\overline{x_j^2}] \simeq \log(\overline{x_m^2}-E[\overline{n_m^2}]) \quad (16)$$

This result means, according to the model for additive noise, that the expected value of the hidden information $\log(\overline{s_m^2})$ is equal to the log of the SS estimation if SS is defined as being $\overline{x_m^2}-E[\overline{n_m^2}]$. If the gain introduced by the transmission channel is considered constant inside each one of the 14 DFT Mel filters the convolutional distortion can be represented by $H=[h_1,h_2,h_3,...,h_m,...h_{14}]$ and due to the fact that H is constant along time

$$E[\log(h_m\cdot\overline{s_m^2})|\overline{x_j^2}] \simeq E[\log(\overline{s_m^2})|\overline{x_j^2}]+h_m^l \quad (17)$$

where $h_m^l=\log(h_m)$. Therefore, the convolutional distortion could be effectively removed after the additive noise being cancelled by means of SS.

5. SS AND MEAN NORMALIZATION

If there is only convolutional distortion, a widely used technique is Cepstral Mean Normalization (CMN). However, when the speech signal is also corrupted by additive signals, CMN loses its effectiveness [2]. Nevertheless, as was shown in the last section, the effect of an unmatched transmission channel could effectively be removed after the additive noise being removed by means of SS given that the SS estimation, $Est(\overline{s_m^2})$, is defined as being equal to $\overline{x_m^2}-E[\overline{n_m^2}]$. Due to the fact that $Est(\overline{s_m^2})$ may be negative in those channels with low SNR a rectifying function $r()$is applied. In order to model the effect introduced by this rectifying function, the distribution of $\overline{n_m^2}$ needs to be known but this is difficult to achieve in real applications where the noise should be estimated in short non-speech intervals.

The insertion of a transmission channel results in an additive constant in both the logarithmic and cepstral domain, and can be cancelled by subtracting the mean from all input vectors. In this paper the mean normalization technique was applied in the logarithmic domain, before the cepstral transform. The mean was computed by

$$\overline{\log[Est(\overline{s_m^2})]} = \frac{\sum_{k=1}^{K}w(k,m)\cdot\log[Est(\overline{s_m^2})]}{\sum_{k=1}^{K}w(k,m)} \quad (18)$$

where K is the number of frames of the utterance (or set of utterances) and $w(k,m)=1$ for the ordinary arithmetic mean. A weighted arithmetic mean was also tested where $w(k,m)$ was defined as:

$$w_{k,m} = \begin{cases} 1 & \text{if } Var[\log(\overline{s_{k,m}^2})|\overline{x_m^2}]\le\delta \\ \frac{\delta}{Var[\log(\overline{s_{k,m}^2})|\overline{x_m^2}]} & \text{if } Var[\log(\overline{s_{k,m}^2})|\overline{x_m^2}]>\delta \end{cases} \quad (19)$$

where $Var[\log(\overline{s_{k,m}^2})|\overline{x_m^2}]$ was estimated according to (10). The idea of (19) is to give a low weight to those bands with low SNR in the computation of the means in order to reduce the effect introduced by the rectifying function.

6. 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) and the car noise from the Noisex database [3]. The isolated words were manually rather than automatically end detected in order to eliminate any effect introduced by the discriminative selection of speech intervals with higher energies. The signal processing was as in [1]. At the output of each channel the energy was computed, SS was applied and the log of the energy was calculated. In every frame, the log energies were normalized to the highest component and 10 cepstral coefficients were computed. In these experiments the noise estimation was made only once using just 250ms of non-speech signal and was kept constant for all the experiments at the same global SNR. The results presented in this paper were achieved with 1000 recognition tests for each SNR. Unless the opposite is specified, a one-step weighted DP algorithm previously proposed in [1] was used in all the experiments. Where spectral tilt experiments were performed clean reference utterances are compared with noisy testing utterances corrupted by additive plus convolutional noise. The tilt applied was a flat frequency response up to a break point frequency of 250Hz followed by a +6dB/oct tilt above 250Hz. The +6dB/oct spectral tilt was chosen instead of +3dB/oct, usually used in many papers, to make the testing conditions more severe.

In all the experiments SS was applied in the linear domain utterance by utterance and the convolutional distortion was cancelled after SS using one, two, five or 10 additive-noise-free utterances (from different words of the

vocabulary) every time by means of mean normalization in the logarithmic domain (LMN). In all the tests where the weighted DP algorithm was used the parameter δ was made equal to 10 in (12) and (19), a value that was shown to be suitable according to some experiments.

In experiments with SS and mean normalization, the means were computed in the logarithmic domain, before cepstral transform. The folowing configurations were tested: SS, SS with ordinary DTW; WSS, SS with the one-step weighted algorithm [1]; $WSS-LMN$, SS and mean normalization with the ordinary arithmetic mean (18); and $WSS-WLMN$, SS and mean normalization with the weighted arithmetic mean (18)(19). The results are shown in Table 1 (without spectral tilt) and Table 2 (with spectral tilt). Figure 1 shows the recognition for $WSS-WLMN$ using different number of utterances to cancel the convolutional noise. In Tables 1 and 2, the means were estimated using 10 additive-noise-free utterances (one per word of the vocabulary.

Table 1: Recognition error rate (%) for speech signal corrupted only by additive noise (car).

SNR	*18dB*	*12dB*	*6dB*	*0dB*
SS	4.4	7.3	12.9	21.5
WSS	0.1	0.4	1.8	8.6
WSS-LMN	0.3	2.9	8.6	28.7
WSS-WLMN	0.3	0.7	2.7	8.2

Table 2: Recognition error rate (%) for speech signal corrupted by additive noise (car) and spectral tilt (6dB/oct).

SNR	*18dB*	*12dB*	*6dB*	*0dB*
SS	24.6	24.8	29.15	36.4
WSS	23.0	26.3	32.5	40.2
WSS-LMN	0.3	1.8	6.9	21.8
WSS-WLMN	0.4	0.7	3.6	10.7

7. DISCUSSION

As can be seen in table 1, WSS (weighted DP algorithm with SS) showed a substantial reduction in the error rate in all the SNR's when compared with SS (ordinary DTW with SS). When compared with WSS, $WSS-LMN$ increased the error rate. However, when the weighted arithmetic mean was used $WSS-WLMN$, the mean normalization almost did not affect the recognition accuracy. According to table 2, the spectral tilt dramatically decreased the recognition accuracy at all the SNR's for SS and WSS. The use of the ordinary mean normalization technique $WSS-LMN$substantially reduced the error rate, but the best results were achieved in $WSS-WLMN$ with the weighted mean. Comparing the results of the table 2 with the ones in table 1, $WSS-WLMN$ was almost completely robust to the convolutional distortion

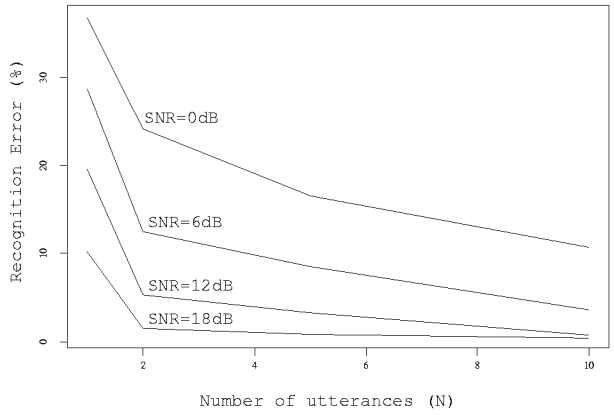

Figure 1: Recognition error rate (%) for speech signal corrupted by additive noise (SNR equal to 18, 12, 6 and 0 dB), and spectral tilt (6dB/oct) as a function of the number of utterances (N) used to estimate the weighted arithmetic mean in $WSS-WLMN$.

at all the SNR's. However, as can be seen in Figure 1, the mean normalization technique is strongly dependent on the length of the speech signal used to estimate the coefficient means and and the required number of utterances apparently increases for lower SNR's.

8. CONCLUSION

The results presented in this paper show that the channel response can effectively be removed after the additive noise being cancelled by means of SS, even when additive noise is estimated with just a few frames. In these experiments the noise estimation was made only once using just 250ms of non-speech signal and was kept constant for all the experiments at the same SNR. Moreover, weighting the information along the noisy speech signal helped to cancel both additive and convolutional noises and good results were achieved with techniques easily implemented such as SS and mean normalization.

References

[1] N.B.Yoma, F.R.McInnes, M.A.Jack. *Weighted Matching Algorithms and Reliability in Noise Cancelling by Spectral Subtraction.* Proceedings ICASSP 97, Vol.2, pp. 1171-1174.

[2] M.F Gales, S.J. Young. *Robust speech recognition in additive and convolutional noise using parallel model combination.* Computer Speech and Language (1995)9, pg. 289-307.

[3] 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.