WEIGHTED VITERBI ALGORITHM AND STATE DURATION MODELLING FOR SPEECH RECOGNITION IN NOISE

*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.
nestor@ccir.ed.ac.uk

ABSTRACT

A weighted Viterbi algorithm (HMM) is proposed and applied in combination with spectral subtraction and Cepstral Mean Normalization to cancel both additive and convolutional noises in speech recognition. The weighted Viterbi approach is compared and used in combination with state duration modelling. The results presented in this paper show that a proper weight on the information provided by static parameters can substantially reduce the error rate, and that the weighting procedure improves better the robustness of the Viterbi algorithm than the introduction of temporal constraints with a low computational load. Finally, it is shown that the weighted Viterbi algorithm in combination with temporal constraints leads to a high recognition accuracy at moderate SNR's without the need of an accurate noise model.

1. INTRODUCTION

In previous papers [1] [2] it was shown that weighting the information along the signal can substantially improve the recognition accuracy when the speech signal is corrupted by additive and convolutional noises, using spectral subtraction (SS) and mean normalization, two easily implemented techniques. In [1] [2] the experiments were done using a one-step weighted DTW and the noise energy in the filter bank was poorly estimated only once using 200 ms of non-speech signal. Those results revealed that the classical concept of matching algorithm where all the frames have the same weight should be revised in order to take into consideration the reliability in noise cancelling frame by frame. It was shown that once the noise is added, an uncertainty is introduced and the original signal cannot be recovered with 100 % accuracy, and the reliability (inverse of uncertainty) in noise cancelling is dependent on the segmental SNR. It is worth noting that reliability weighting could be considered as a formalization of a very important characteristic of the auditory perception which does not have to recover all the information of the corrupted speech signal and reduces the importance of the more noisy intervals to extract the information that is relevant to understand the message.

The contributions of this paper concern: a) a weighted Viterbi algorithm and a weighting function without a free variable; b) combination of this modified Viterbi algorithm with SS and Cepstral Mean Normalization (CMN), two easily implemented techniques; c) comparison and combination of the weighted Viterbi algorithm with state duration modelling. It is shown that weighting the information along the signal requires a low computational load and leads to better results than the introduction of temporal constraints in the recognition algorithm. In combination with temporal constraints, the weighted Viterbi algorithm resulted in a high recognition accuracy at SNR=18, 12 and 6dB without a noise model. 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 authors believe that weighted matching algorithm approach could be applied to other problems of robust processing such as speaker verification and speaker adaptation in noisy conditions.

*Supported by a grant from CNPq-Brasilia/Brasil

2. RELIABILITY IN NOISE CANCELLING

In [1] [2] it was suggested that the hidden clean information of the speech signal is a function of the observed noisy signal energy $\overline{x_m^2}$, the noise energy $\overline{n_m^2}$ and the phase difference ϕ_m between the clean signal and noise in channel m:

$$\overline{s_m^2}(\phi_m, \overline{n_m^2}, \overline{x_m^2}) = 2 \cdot A^2 \cdot cos^2(\phi_m) + B - 2 \cdot A \cdot cos(\phi_m) \cdot \sqrt{A^2 \cdot cos^2(\phi_m) + B} \quad (1)$$

where $A = \sqrt{\overline{n_m^2} c_m}$, $B = \overline{x_m^2} - \overline{n_m^2}$ and c_m is a correction coefficient. Using (1) and assuming that the random variables ϕ_m and $\overline{n_m^2}$ are uncorrelated, ϕ_m is uniformly distributed between $-\pi$ and π and that $\overline{n_m^2}$ is concentrated near its mean $E[\overline{n_m^2}]$, it is possible to estimate [2]

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

by means of replacing B with

$$\overline{B} = max\left\{(\overline{x_m^2} - E[\overline{n_m^2}]), SsThr_m\right\}$$

where $SsThr_m$ is a constant. This rectifying function is needed because $\overline{x_m^2} - E[\overline{n_m^2}]$ may be negative due to ϕ_m and the variation of the noise.

The variance $Var[\log(\overline{s_m^2})|\overline{x_m^2}]$ is an estimation of the uncertainty related to noise cancelling and was used to weight the matching algorithm [1] [2] and it was proved, by means of a modified version of the DTW algorithm, that weighting the information along the signal could substantially reduce the error rate when the clean signal was corrupted by additive and convolutional noises using a poor estimation of noise. The frame weighting function was defined as being

$$w = \begin{cases} 1 & \text{if } TotalVar \leq VarThr \\ \frac{VarThr}{TotalVar} & \text{if } TotalVar > VarThr \end{cases} \quad (3)$$

where

$$TotalVar \quad = \quad \sum_{m=1}^{14} Var[\log(\overline{s_m^2})|\overline{x_m^2}] \qquad (4)$$

If $SsThr_m$ is low when compared to the noise estimation $E[\overline{n_m^2}]$, $Var[\log(\overline{s_m^2})|\overline{x_m^2}]$ may be too high at low segmental SNR's (when the model loses accuracy). This is counteracted by setting an upper bound to $Var[\log(\overline{s_m^2})|\overline{x_m^2}]$ equal to $Var[\log(\overline{s_m^2})]$ which is estimated on the clean signal.

The function represented by (3) sets that the frame weight would be inversely proportional to the sum of the uncertainty variances in all the DFT Mel filters, and $VarThr$ is a threshold introduced due to the fact that $TotalVar$ is theoretically equal to zero for clean speech signal. In the context of the weighted DTW [1] [2], (3) strongly reduced the error rate in all the SNR's but the optimal threshold $TotalVar$ was case dependent, although it presented a wide range of only slightly suboptimal values.

3. WEIGHTED VITERBI ALGORITHM

The reliability coefficient can be included in the Viterbi algorithm [3] by raising the output probability of observing the frame T_t to the power of $w(t)$, where t is the time index. This modification leads to the following algorithm:

STEP 1 : Initialization. For each state i,

$$\delta_1(i) = \pi_i \times [b_i(T_1)]^{w(1)}$$

$$\psi_1(i) = 0$$

STEP 2 : Recursion. From t=2 to L_T, for all states j,

$$\delta_t(j) = Max_i[\delta_{t-1}(i) \times a_{ij}] \times [b_j(T_t)]^{w(t)}$$

$$\psi_t(j) = argmax_i[\delta_{t-1}(i) \times a_{ij}]$$

STEP 3: End: (* indicates the optimised results).

$$P^* = Max_{s \in Sf}[\delta_{L_T}(s)]$$

where L_T is the frame sequence length and s_F is the set of possible final states. Consequently, the influence of the probability $b_i(T_{t-1})$ in the decision $Max_i[\delta_{t-1}(i) \times a_{ij}] = Max_i[Max_h[\delta_{t-2}(h) \times a_{hi}] \times [b_i(T_{t-1})]^{w(t-1)} \times a_{ij}]$ at STEP 2 depends on $w(t-1)$: if $w(t-1) = 1$ (high reliability), the influence of $b_i(T_{t-1})$ is maximum; if $w(t-1) = 0$ (very low reliability), the influence of $b_i(T_{t-1})$ is zero because $[b_i(T_{t-1})]^0 = 1$ for all states i.

Preliminary experiments with the modified version of the Viterbi algorithm were done using (3) where $TotalVar$ was computed using the uncertainty variances in the logarithm domain as defined in (2). Results mainly confirmed the previous experiments with the weighted DP equation , but another weighting function was defined in the cepstral domain using the HMM variances in order to eliminate the threshold $VarThr$

3.1. Mapping from the Log to the Cepstral Domain

The cepstral coefficients are estimated by means of

$$c_n \quad = \quad \sum_{m=1}^{M} E_m \cdot cos[\frac{\pi \cdot n}{N} \cdot (m - 0.5)] \qquad (5)$$

where E_m is the logarithm of the energy at the output of the filter m that results from the SS estimation, and N is the number of cepstral coefficients. If the log energies are supposed uncorrelated, $Var[c_n|X]$ can be re-written as

$$Var[c_n|X] \quad = \quad \sum_{m=1}^{M} Var[\log(\overline{s_m^2})|\overline{x_m^2}] \cdot cos^2[\frac{\pi n}{N}(m - 0.5)] \qquad (6)$$

The components $\overline{s_m^2}(\phi_m, \overline{n_m^2}, \overline{x_m^2})$ depend on the clean speech and noise signals and are clearly correlated specially for contiguous filters. However, although the uncorrelated condition was a rough approximation, it was enough to lead to good results.

3.2. Modified weighting function

The weighting function that (3) attempts to model could also be approximated by

$$w \quad = \quad \frac{VarThr}{VarThr + TotalVar} \qquad (7)$$

Experiments with the weighted version of DTW showed that (3) and (7) lead to similar results. In order to avoid the threshold $VarThr$, a weighting function based on (7) was proposed using the variances of the HMM's. In the experiments here reported, each word was modelled using an 8-state left-to-right topology without skip-state transition, with a single multivariate Gaussian density per state and a diagonal covariance matrix, and the modified frame weighting function is defined as

$$w \quad = \quad \frac{1}{N} \sum_{n=1}^{N} \frac{\sigma_{\lambda,i,n}^2}{\sigma_{\lambda,i,n}^2 + Var[c_n|X]} \qquad (8)$$

where $\sigma_{\lambda,i,n}^2$ is the variance of coefficient n, state i and model λ. The function shown in (8) compares the uncertainty variance of coefficient n with the variance of the coefficient n in a phonetic class or state of a HMM. Moreover, if uncertainty variance is high for one coefficient, w is not necessarily low because the weight is the sum of terms $\frac{\sigma_{\lambda,i,n}}{\sigma_{\lambda,i,n} + Var[c_n|X]}$. Finally, if the signal is clean $Var[c_n|X]$ is zero for all n and $w = 1$.

4. TEMPORAL CONSTRAINTS

In the ordinary HMM topology, the transition probability is represented by a constant that leads to a geometric probability density for state duration which is not accurate for most cases. Several methods to include temporal constraint have been proposed. Parametric state duration distributions, Poisson [4] and gamma [5], were used in the training process but the method requires a high computational load. In [6] it was proposed a backtracking procedure where the duration contribution to the standard Viterbi metric is added after collecting possible candidate paths. The disadvantage of this approach is that the correct alignment path may not be one of these candidates. A significant improvement of the error rate when the speech signal was corrupted by additive noise was reported in [7] by means of introducing the state duration constraints in the training procedure, using the state sequences that are likely to happen and fulfill the temporal restrictions.

In order to include temporal constraints in the HMM recognizer, it was followed the procedure suggested by [8] where the the state durations are modelled using gamma distributions. Every state was associated to a gamma distribution whose parameters were estimated using the training database after the HMM's had been trained. The discrete gamma distribution is given by [8]:

$$d(\tau) \quad = \quad K \cdot e^{-\alpha \cdot \tau} \cdot \tau^{p-1} \tag{9}$$

where $\tau = 0, 1, 2, ...$ is the duration of a given state in number of frames, $\alpha > 0$, $p > 0$ and K is a normalizing term. This distribution was proved to fit better the empirical (state and word) duration distributions than the Gaussian or geometric functions [8]. After training the HMM's, the optimal state sequence was estimated for every training utterace using the Viterbi algorithm and the parameters α and p were estimated for every state in each model by means of:

$$\alpha \quad = \quad \frac{E(\tau)}{Var(\tau)} \tag{10}$$

$$p \quad = \quad \frac{E^2(\tau)}{Var(\tau)} \tag{11}$$

where $E(\tau)$ and $Var(\tau)$ are, respectively, the mean and variance of the state duration directly computed using Viterbi alignment. Beside $E(\tau)$ and $Var(\tau)$, $min(\tau)$ and $max(\tau)$ were also estimated.

Instead of using the duration metric suggested in [8], the transition probabilities were defined as

$$a_{i,i}^{(\tau)} = Prob(s_{t+1} = i | s_t = s_{t-\tau+1} = i)$$

$$a_{i,j}^{(\tau)} = Prob(s_{t+1} = j | s_t = s_{t-1} = ... = s_{t-\tau+1} = i)$$

Using these definitions for the transition probabilities, $a_{i,i}^{(\tau)}$ and $a_{i,j}^{(\tau)}$ can be estimated by

$$a_{i,i}^{(\tau)} \quad = \quad \frac{D_i(\tau) - d_i(\tau)}{D_i(\tau)} \tag{12}$$

$$a_{i,j}^{(\tau)} \quad = \quad \frac{d_i(\tau)}{D_i(\tau)} \tag{13}$$

where $D_i(\tau)$ is the probability of state i being active for $t \geq \tau$:

$$D_i(\tau) \quad = \quad \sum_{t=\tau}^{t_{max}} d(t) \tag{14}$$

In order to include the possible min and max durations, the transition probabilities were modified to:

$$a_{i,i} \quad = \quad \begin{cases} 1 & \text{if } \tau < t_{min} \\ 0 & \text{if } \tau \geq t_{max} \\ a_{i,i}^{(\tau)} & \text{otherwise} \end{cases} \tag{15}$$

$$a_{i,j} \quad = \quad \begin{cases} 0 & \text{if } \tau < t_{min} \\ 1 & \text{if } \tau \geq t_{max} \\ a_{i,j}^{(\tau)} & \text{otherwise} \end{cases} \tag{16}$$

where $t_{min} = 0.8 \cdot min(\tau)$ and $t_{max} = 1.5 \cdot max(\tau)$. The constants 0.8 and 1.5 introduce a tolerance to the min and max duration for every state.

The recognition experiments were speaker dependent using isolated words (digits). In some cases it was observed that the variation in state duration was very low, which resulted in a low $Var(\tau)$ which in turn caused a low recognition accuracy. To counteract this, a threshold was introduced to set a floor for $Var(\tau)$. According to some experiments, a suitable value for this threshold would be 4.

5. 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 and speech noises from the Noisex database [9]. Where convolutional noise experiments were performed a spectral tilt composed by a flat frequency response up to a break point frequency of 200Hz followed by a +3dB/oct tilt above 250Hz was applied to the noisy signals. The signals were downsampled to 8000 samples/sec. The 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 DFT filters. At the output of each channel the energy was computed, SS [10] and CMN were applied and the log of the energy was estimated. The overestimation parameter for SS $\alpha = 2.0$ and the noise spectral floor $\beta = 0.01$. In every frame 10 cepstral coefficients were computed. In tests with only additive noise, only SS was used. In tests with both additive and convolutional noises, CMN was applied after SS and the coefficient means were computed using one utterance per word of the vocabulary (digits) every time.

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 threshold $SsThr_m$ (used to compute $\overline{B}$ in (2)) was estimated according to [11] and was approximately equal to 20dB for all the channels. Each word was modelled using an 8-state left-to-right HMM without skip-state transition, with a single multivariate Gaussian density per state and a diagonal covariance matrix. The HMM's and the state duration distributions were estimated by means of the clean signal training utterances. In the experiments HTK V.2.0 with modifications to include the temporal constraints and reliability weighting in the testing procedure was used for the HMM experiments. The following configurations were tested: the ordinary Viterbi algorithm Vit ; the weighted version of the Viterbi algorithm using (8) as the weighting function, $W2 - Vit$; the ordinary Viterbi algorithm with state duration constraints using gamma distributions, $Vit-T$; and finally, $W1-Vit-T$ and $W2-Vit-T$, the weighted algorithm with temporal constraints using, respectively, (3) and (8) as weighting coefficients.

6. DISCUSSION AND CONCLUSION

As can be seen in Tables 1 and 2, the weighted version of the Viterbi algorithm using (8), $W2 - Vit$, as weighting function strongly reduced the error rate at all the SNR's. The ordinary Viterbi algorithm with temporal constraints $Vit - T$ also reduced the error rate but the improvement was poorer than with the weighted algorithm. The best results were achieved when weighting procedure was applied in combination with state duration modelling $W2-Vit-T$.

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

SNR	*18dB*	*12dB*	*6dB*	*0dB*
Vit	1.5	16.5	66	86
W2-Vit	0	0	2	26
Vit-T	0	5	15.5	28.5
W2-Vit-T	0	0	0	13

Table 2. Recognition error rate(%) for speech corrupted by additive noise (speech noise).

SNR	*18dB*	*12dB*	*6dB*	*0dB*
Vit	4.5	38.5	78.5	90
W2-Vit	0	2	12	58
Vit-T	1	4.5	14.5	38
W2-Vit-T	0	0	4.5	38

It is interesting to highlight that weighting the information along the signal requires a low computational load and was more effective than the introduction of the temporal constraints. In other words, the weighted Viterbi algorithm is more robust to unlikely alignments because the recognition tends always to rely on those frames with higher segmental SNR. Although not shown in Tables 1 and 2, the weighted algorithm using (3) gave better results than $Vit-T$ but worse than $W2-Vit$. According to Fig. 1, the weighting function (8) $W2-Vit-T$ led to better results than the ordinary Viterbi algorithm just with temporal constraints $Vit-T$ and than $W1-Vit-T$ (using (3) and with the optimal $VarThr$) without the need of a free variable. Although not reported in this paper, this behaviour was also observed for all the noises from [9] considered in this research (car, speech, Lynx, oper.room and factory).

Tests with additive and convolutional noise (Tables 3 and 4) indicate that the weighting procedure can also be effective if CMN is applied after SS. The weighted Viterbi algorithm is only one step and can be used by either isolated or continuous word recognition and a high accuracy was observed at SNR= 18, 12 and 6dB using no free variables (except the ones related to SS), only static parameters and a poor estimation of noise made in just 200ms. Currently work is being done to improve the accuracy at lower SNR. Finally, it is worth mentioning that weighted matching algorithms could also be used with other noise cancelling methods.

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] N.B.Yoma, F.R.McInnes, M.A.Jack. *Spectral Subtraction and Mean Normalization in the context of weighted matching algorithms.* Proceedings Eurospeech 97, Vol.3, pp. 1411-1414.

Table 3. Recognition error rate(%) for additive (car noise) and convolutional (3dB/oct) noises.

SNR	*18dB*	*12dB*	*6dB*	*0dB*
Vit-T	1.5	5	14.5	39
W2-Vit-T	0.0	0	0.5	22

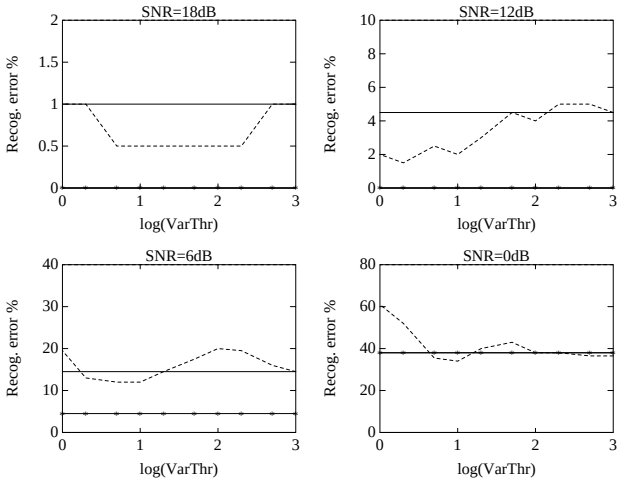

Figure 1. Recognition error rate(%) for speech signal corrupted by additive noise (speech noise): (–), *Vit-T*; (- -), *W1-Vit-T*; and (-*-), *W2-Vit-T*.

Table 4. Recognition error rate(%) for additive (speech noise) and convolutional (3dB/oct) noises.

SNR	*18dB*	*12dB*	*6dB*	*0dB*
Vit-T	4.5	6	23	44
W2-Vit-T	0.0	0.5	4.5	32

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

[4] M.J.Russell, R.K.Moore. *Explicit Modelling of State Occupancy in Hidden Markov Models for Automatic Speech Recognition.* Proc. ICASSP 1985, pp. 5-8.

[5] S.E.Levinson. *Continuously Variable Duration Hidden Markov Models for Automatic Speech Recognition.* Computer Speech and Language, Vol. 1, pp. 29-45, 1986.

[6] L.R. Rabiner, J.G. Wilpon, and F.K. Soong. *High Performance connected digit recognition using Hidden Markov Models.* IEEE Trans. ASSP, vol.37, pp. 1214-1225, Aug.1989.

[7] K. Laurila. *Noise robust speech recognition with state duration constraints.* Proc. ICASSP 97, Vol.2, pp.871-874

[8] D. Burshtein. *Robust Parametric Modeling of Durations in Hidden Markov Models* . IEEE Trans. ASSP, vol.4, No. 3, May 1996.

[9] 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.

[10] M.Berouti, R.Schwartz, J.Makhoul. *Enhancement of Speech Corrupted by Acoustic Noise.* Proc. ICASSP, pp.208-211, 1979.

[11] D. Van Compernolle. *Noise adaptation in a hidden Markov model speech recognition system.* Computer Speech and Language (1989)3, pp.151-167.