# A Noise-Robust ASR Back-end Technique Based on Weighted Viterbi Recognition

*Xiaodong Cui, Alexis Bernard* and  Abeer Alwan*

Department of Electrical Engineering
University of California, Los Angeles, CA
Email: {xdcui, abernard, alwan}@icsl.ucla.edu

## Abstract

The performance of speech recognition systems trained in quiet degrades significantly under noisy conditions. To address this problem, a Weighted Viterbi Recognition (WVR) algorithm that is a function of the SNR of each speech frame is proposed. Acoustic models trained on clean data, and the acoustic front-end features are kept unchanged in this approach. Instead, a confidence/robustness factor is assigned to the output observation probability of each speech frame according to its SNR estimate during the Viterbi decoding stage. Comparative experiments are conducted with Weighted Viterbi Recognition with different front-end features such as MFCC, LPCC and PLP. Results show consistent improvements with all three feature vectors. For a reasonable size of adaptation data, WVR outperforms environment adaptation using MLLR.

## 1. Introduction

Noise-robust speech recognition is an important challenge for real world applications. The performance of recognition systems trained in quiet degrades significantly in the presence of background acoustic noise. In general, there are two ways of addressing this problem. The first approach is to reduce mismatch in the front end feature extraction stage [1] [2]. The other approach involves either updating 'clean' acoustic models based on noise estimates [3] or building separate HMMs of the 'clean' speech and of the noise [4].

In [7], a Weighted Viterbi Recognition (WVR) algorithm was introduced to deal with channel impairments, frame erasures and network congestion for Distributed Speech Recognition (DSR). Also, independent work was conducted in [9] using "soft-feature" decoding to deal with DSR channel degradation. In this paper, we use the WVR algorithm to deal with background acoustic noise without changing the acoustic speech models.

---

The weighting factor is a function of the SNR estimate of each speech frame. The computational complexity of this algorithm is quite low and its structure renders it easy to implement in DSR systems. Compared with environment adaptation using MLLR with a reasonable size of adaptation data, WVR can achieve better results. Three types of feature vectors are examined: MFCC, LPCC and PLP [2].

The remainder of this paper is organized as follows. In Section 2, a system overview is provided. In Sections 3 and 4, the SNR estimation algorithm and WVR formulation are described, respectively. Experimental results are shown in Section 5, and Section 6 concludes the paper with a summary and discussion.

## 2. System Overview

A system overview is illustrated in Fig. 1, where acoustic HMMs are trained using clean data and front-end feature extraction using standard features such as MFCC, LPCC and PLP. The SNR is estimated for each speech frame and the estimate is provided to a Viterbi decoding/recognition module where a final decision is made based on the clean acoustic models and the confidence/quality of each speech frame.
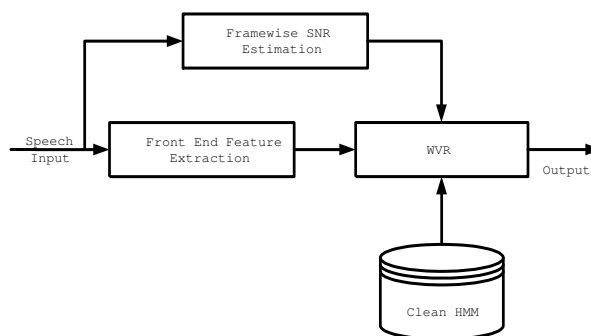


Figure 1: Weighted Viterbi Recognition (WVR) to deal with noisy speech given 'clean' acoustic models.

## 3. Frame-based SNR Estimation

The weighting factor in WVR is a function of SNR estimates. In this paper, we refer to the average utterance signal-to-noise ratio, and the frame-based signal-to-noise ratio as $SNR_{av}$ and SNR, respectively.

A minimum statistics tracking method is adopted ([5] and [6]). Assuming that the noisy speech power is the summation of power of clean speech and background noise, tracking power spectral minima can provide fairly accurate estimation of the background noise power, hence good estimation of SNR. Also, by tracking minimum statistics, this algorithm can deal with nonstationary background noise with slowly changing statistical characteristics. One disadvantage of this approach is the bias between the mean and minimum value of the background noise. Hence, in this paper, a constant factor is applied to compensate for the bias. Power spectral minimum statistics is searched within a 0.5 second interval preceding each speech frame. In real applications, this will introduce an extra memory requirement and time delay. But, compared with other complicated front end processing algorithms, the overhead is quite small. Figs. 2 and 3 show estimated frame-based SNRs and the confidence factor ($\gamma_t$) for two speech utterances from the Aurora 2 database. In Aurora 2, the signal-to-noise ratio is estimated for the whole utterance. From these figures we can see that frame-based SNR values vary in a large range compared to the overall utterance SNR. Also notice that there is an SNR floor set at 0 dB for all frames because we assume that SNR estimates below 0 dB are not reliable.
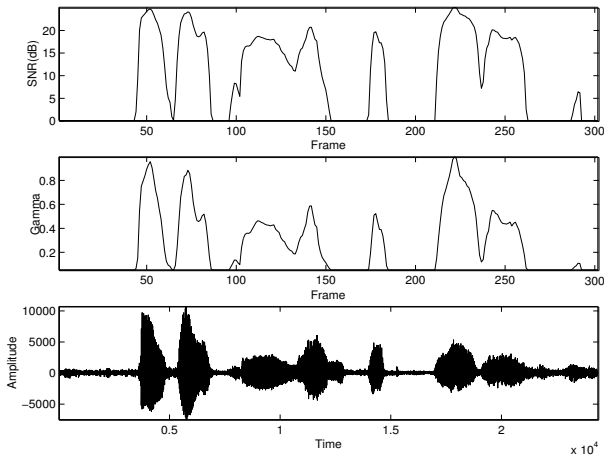


Figure 2: Estimated SNR and confidence factor ($\gamma_t$) for the utterance "0021641" labeled in the Aurora 2 database as having a signal-to-noise ratio of 15 dB.

## 4. Weighted Viterbi Recognition

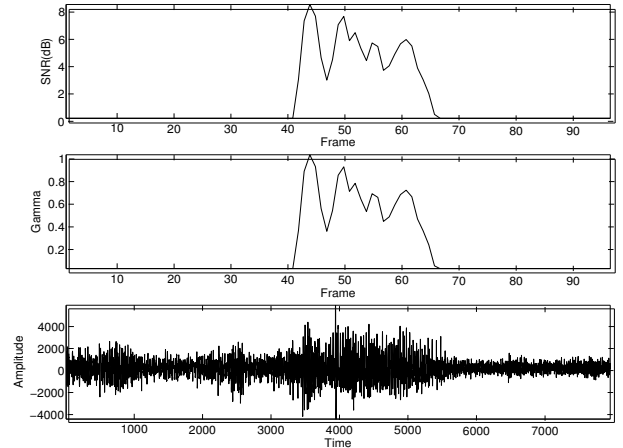WVR modifies the recursive step of the Viterbi algorithm to take into account the effect of SNR by weighting the



Figure 3: Estimated SNR and confidence factor ($\gamma_t$) for the utterance "4" labeled in the Aurora 2 database as having a signal-to-noise ratio of 0 dB.

probability of observing features given the HMM state model $b_j(o_t)$ with the confidence factor of current feature observation $o_t$. The time-varying confidence factor $\gamma_t$ can be inserted into the Viterbi algorithm by raising the probability $b_j(o_t)$ to the power $\gamma_t$ to obtain the following state update equation [7]:

$$\phi_j(t) = \max_i \{\phi_i(t-1) \cdot a_{ij}\}[b_j(o_t)]^{\gamma_t} \qquad (1)$$

where $\phi_j(t)$ represents the maximum likelihood of observing speech feature $o_1$ to $o_t$ and being in state $j$ at time $t$, $a_{ij}$ stands for the transition probability from state $i$ to state $j$ and $\gamma_t \in [0,1]$ is a time-varying frame confidence factor that maps the frame SNR into the interval $[0,1]$. The value(s) of $\gamma_t$ are determined empirically.

In extreme cases, when $\gamma_t = 0$, $\phi_j(t)$ are updated only by state transition probability $a_{ij}$ and the probability $b_j(o_t)$ for current frame is discarded; when $\gamma_t = 1$, current frame is decoded by regular unweighted Viterbi recognition scheme.

In this paper, two sets of $\gamma_t$ are chosen depending on the average signal-to-noise ratio of the utterance $SNR_{av}$ as shown in Fig. 4.

When the utterance $SNR_{av}$ is above 10 dB, then

$$\gamma_t = \begin{cases} 1 & \text{for SNR} \geq 25 \text{ dB} \\ e^{0.12 \cdot (SNR-25)} & \text{for SNR} < 25 \text{ dB}. \end{cases} \qquad (2)$$

If the utterance $SNR_{av}$ is less than 10 dB, a simple normalization by the maximum SNR value of the utterance is adopted:

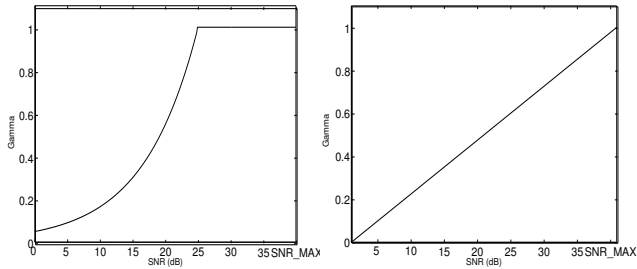$$\gamma_t = \frac{SNR}{SNR_{max}} \qquad (3)$$

Figure 4: $\gamma_t$ (gamma) for $SNR_{av}$ higher than 10 dB (left) and $\gamma_t$ (gamma) for $SNR_{av}$ lower than 10 dB (right).

## 5. Experimental Results

### 5.1. WVR with MFCC, LPCC and PLP

Comparative experiments are carried out for the WVR algorithm using different features: MFCC, LPCC and PLP. The experiments use Aurora 2 database. Training corpus has only clean speech data while the test corpus contains noisy speech at signal-to-noise levels labeled as clean, 20 dB, 15 dB, 10 dB, 5 dB and 0 dB. Left-to-right topology is adopted for the HMMs. Each digit has 16 states with 3 mixtures each. There are one 3-state silence model with 6 mixtures and one short pause model who shares the middle state of the silence model. The above topology is the standard setup provided by the Aurora 2 database [8].

Tables 1, 2 and 3 show the WVR performance improvement over baseline performance for an MFCC, LPCC and PLP frontend. In test corpus, Set A and Set B are chosen from the Aurora 2 database since only additive noise is tested in this paper. In each set, there are four types of noise with different frequency characteristics which are subway, babble, car and exhibition noise for Set A, and restaurant, street, airport and station noise for Set B. For each type of noise, six noise levels are tested ranging from clean to 0 dB as shown in the tables. For each noise level, word accuracy averaged over all the four noise types are presented.

| | Set A | | | Set B | | |
|---|---|---|---|---|---|---|
| | Baseline | WVR | Imprv.(%) | Baseline | WVR | Imprv.(%) |
| Clean | 98.94 | 98.96 | 1.9 | 98.94 | 98.96 | 1.9 |
| 20 dB | 94.99 | 96.50 | 30.1 | 92.35 | 96.42 | 53.2 |
| 15 dB | 86.93 | 92.89 | 45.6 | 80.79 | 91.83 | 57.5 |
| 10 dB | 67.28 | 84.70 | 53.2 | 58.06 | 86.46 | 67.7 |
| 5 dB | 39.36 | 62.99 | 39.0 | 32.04 | 63.46 | 46.2 |
| 0 dB | 17.07 | 34.91 | 21.5 | 14.63 | 35.30 | 24.2 |

Table 1: WVR performance with MFCC features.

From the tables, we observe that WVR resulted in consistent improvements for all noise conditions and noise levels. On average, the algorithm reduces the error rate by 38%, 45% and 47% for MFCC, LPCC and PLP features, respectively compared with their baselines in

| | Set A | | | Set B | | |
|---|---|---|---|---|---|---|
| | Baseline | WVR | Imprv.(%) | Baseline | WVR | Imprv.(%) |
| Clean | 98.71 | 98.73 | 1.6 | 98.71 | 98.73 | 1.6 |
| 20 dB | 91.09 | 94.89 | 42.7 | 86.77 | 93.88 | 53.7 |
| 15 dB | 76.69 | 90.74 | 60.3 | 69.38 | 92.46 | 75.4 |
| 10 dB | 52.86 | 79.43 | 56.4 | 45.88 | 81.64 | 66.1 |
| 5 dB | 27.16 | 57.73 | 42.0 | 24.31 | 59.08 | 45.9 |
| 0 dB | 12.11 | 31.59 | 22.2 | 9.71 | 32.89 | 25.7 |

Table 2: WVR performance with LPCC features.

| | Set A | | | Set B | | |
|---|---|---|---|---|---|---|
| | Baseline | WVR | Imprv.(%) | Baseline | WVR | Imprv.(%) |
| Clean | 98.90 | 99.00 | 9.1 | 98.90 | 99.00 | 9.1 |
| 20 dB | 93.81 | 96.49 | 43.3 | 91.71 | 96.19 | 54.0 |
| 15 dB | 82.13 | 92.35 | 57.2 | 78.04 | 91.04 | 59.2 |
| 10 dB | 59.85 | 84.84 | 62.2 | 53.97 | 86.92 | 71.6 |
| 5 dB | 33.57 | 63.73 | 45.4 | 27.55 | 65.41 | 52.3 |
| 0 dB | 13.35 | 35.92 | 26.1 | 10.36 | 37.15 | 29.9 |

Table 3: WVR performance with PLP features.

Set A and 50%, 53% and 53% for Set B. For both sets of data, the average improvements did not include the clean condition. The highest error reduction is achieved at 10 dB signal-to-noise ratio. Furthermore, WVR performs better for data from Set B compared to Set A. MFCC features give the best baseline results. After WVR, PLP and MFCC have comparable performance.

### 5.2. WVR vs. MLLR

WVR constitutes a very simple back-end technique aiming at improving recognition accuracy under noisy conditions. In this Section, we carry a comparative study between WVR and environment adaptation using Maximum likelihood Linear Regression (MLLR) [3] technique. For MLLR, 40 utterances are randomly selected from each type of noise data set which means the total adaptation data size is 320 sentences.

Fig. 5 illustrates the performance of WVR and MLLR compared with the baseline performance for the Aurora 2 database with 8 types of noise and using MFCC features. For each type of noise, an average over all the signal-to-noise levels (clean, 20 dB, 15 dB, 10 dB, 5 dB and 0 dB) are calculated. Both WVR and MLLR result in improvements over the baseline. Moreover, WVR outperforms MLLR by 2.4% on average without the need for a priori knowledge of noise statistics nor the need for off-line training while MLLR has these requirements.

## 6. Conclusions

In this paper, a Weighted Viterbi Recognition (WVR) algorithm is used in a DSR system to deal with background noise. A confidence factor is assigned to each speech
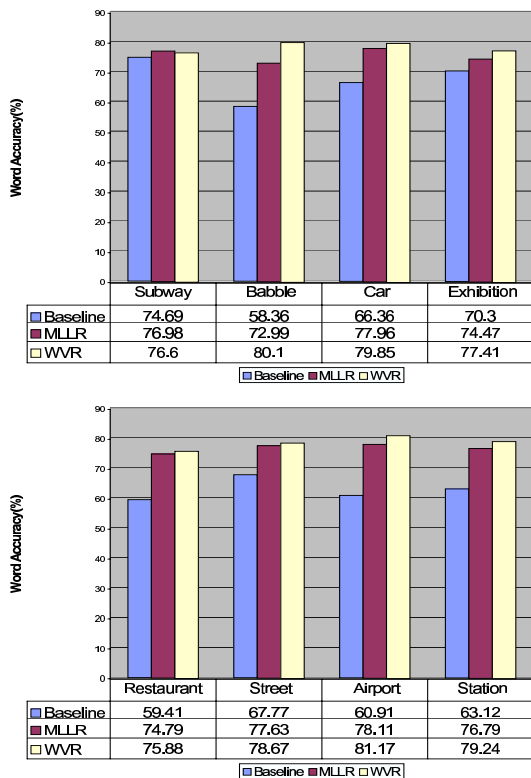
| | Subway | Babble | Car | Exhibition |
|---|---|---|---|---|
| Baseline | 74.69 | 58.36 | 66.36 | 70.3 |
| MLLR | 76.98 | 72.99 | 77.96 | 74.47 |
| WVR | 76.6 | 80.1 | 79.85 | 77.41 |

Baseline ■ MLLR □ WVR



| | Restaurant | Street | Airport | Station |
|---|---|---|---|---|
| Baseline | 59.41 | 67.77 | 60.91 | 63.12 |
| MLLR | 74.79 | 77.63 | 78.11 | 76.79 |
| WVR | 75.88 | 78.67 | 81.17 | 79.24 |

Baseline ■ MLLR □ WVR

Figure 5: Perfomance of WVR vs. MLLR for different types of noise from the Aurora 2 database Set A (top) and Set B (bottom).

frame based on its SNR estimate. The trained acoustic models and the acoustic features are not changed or updated in this approach. Experimental results show consistent improvements of WVR with MFCC, LPCC and PLP features with different types of background noise and SNR levels. For Aurora 2 database, word error rate reduction, compared to baseline performance, was on average 50%. Compared to environment adaptation techniques, WVR outperforms MLLR by 2.4%. WVR has low computational complexity and is easy to incorporate into DSR systems without much memory requirements and time delay.

Note that when calculating $SNR_{av}$ for continuous speech recognition, the values can be updated adaptively based on segmental SNR (e.g. 0.5 secs as in the minimum statistics tracking algorithm), without waiting for the whole utterance to be completed. In this way, time delay can be reduced.

In WVR, weights are assigned to observation features based on frame-based confidence measure; it does not reduce the mismatch between clean models and noisy features. Thus, finding an effective way to reestimate SNR for each frame after mismatch reduction in the feature domain is a promising direction for future work.

## 7. Acknowledgments

## 8. References

[1] X. Cui, M. Iseli, Q. Zhu, and A. Alwan, "Evaluation of Noise Robust Features on the Aurora Databases", *Proc. of ICSLP*, vol.1, Pp.481-484, 2002.

[2] H. Hermansky and N. Morgan, "Rasta Processing of Speech", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol.6, Pp.578-589, 1994.

[3] C. Leggetter and P. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", *Computer Speech and Language*, vol.9, Pp.171-185, 1995.

[4] M. Gales and S. Young, "HMM recognition in Noise Using Parallel Model Combination", *Proc. of Eurospeech*, vol.2, Pp. 873-840, 1993.

[5] R. Martin "An Efficient Algorithm to Estimate Instantanous SNR of Speech Signals", *Proc. of Eurospeech*, vol.3, Pp.1093-1096, 1993.

[6] R. Martin. "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics", *IEEE Trans. on Speech and Audio Processing*, vol.9, No.5, Pp.504-512, July 2001.

[7] A. Bernard and A. Alwan. "Low-bitrate Distributed Speech Recognition for Packet-based and Wireless Communication", *IEEE Trans. on Speech and Audio Processing*, vol.10, No.8, Pp.570-580, Nov. 2002.

[8] H. Hirsch and D. Pearce. "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems Under Noisy Condition", *ISCA ITRW ASR2000 Automatic Speech Recognition: Chanllenges for the Next Millennium*, France, 2000.

[9] A. Potamianos and V. Weerackody, "Soft-feature decoding for speech recognition over wireless channels", *Proc. of ICASSP*, vol.1, Pp. 269-272, 2001.