NOISY SPEECH RECOGNITION USING ROBUST INVERSION OF HIDDEN MARKOV MODELS

Seokyong Moon, *Jenq-Neng Hwang*

Information Processing Laboratory
Department of Electrical Engineering, FT-10
University of Washington, Seattle, WA 98195
e-mail: moon@pierce.ee.washington.edu, hwang@ee.washington.edu

ABSTRACT

The hidden Markov model (HMM) inversion algorithm is proposed and applied to robust speech recognition for general types of mismatched conditions. The Baum-Welch HMM *inversion* algorithm is a dual procedure to the Baum-Welch HMM *reestimation* algorithm, which is the most widely used speech recognition technique. The *forward training* of an HMM, based on the Baum-Welch reestimation, finds the model parameters λ that optimize some criterion, usually maximum likelihood (ML), with given speech inputs s. On the other hand, the *inversion* of an HMM finds speech inputs s that optimize some criterion with given model parameters λ. The performance of the proposed HMM inversion, in conjunction with HMM reestimation, for robust speech recognition under additive noise corruption and microphone mismatch conditions is favorably compared with other noisy speech recognition techniques, such as the projection-based first-order cepstrum normalization (FOCN) and the *robust* minimax (MINIMAX) classification techniques.

1. INTRODUCTION

In the real world, an automatic speech recognition (ASR) system experiences severe performance degradation due to the *mismatch* between training and testing environments. The mismatch between training and testing conditions result from various types of sources, e.g., ambient background noise, microphone mismatch, or various speech styles [1]. Many researchers tried to combat the variety of mismatches by designing robust ASR systems over the recent years. In general, the mismatch compensation methods adopted in most speech recognition systems can be classified into four major categories.

- Compensation before recognition stage.
- Compensation during recognition stage.
- Robust estimation of feature vectors.
- Inclusion of noise statistics to model.

Especially, the second category adjusts the model parameters of speech recognizers (instead of modifying the input speech) so that the models adapt to mismatched conditions, e.g., projection-based first-order cepstral normalization (FOCN) [2] and *robust* minimax (MINIMAX) classification [1]. Since these techniques compensate the noise during the recognition stage, it is relatively easy to incorporate the classification error with appropriate optimization criteria, e.g., minimum classification error (MCE) [3] .

It is empirically and theoretically proved that the norm of LPC cepstral coefficients shrink under the influence of AWGN [2]. The FOCN technique for hidden Markov models (HMMs) compensates the norm shrinkage of LPC cepstral coefficients by simply shrinking the means of Gaussian mixtures of HMM in proportion to the *projection* (or directional cosine) in feature vector space. The performance improvement of FOCN is quite limited due to the very restricted movement of the Gaussian mixture means toward the origin. The MINIMAX technique [1] which utilizes the Baum-Welch reestimation is thus developed to accommodate the HMMs for more general mismatched environments by allowing more flexible movements of the means of Gaussian mixtures toward the noisy speech features with appropriate constraints.

For a continuous density multi mixture (CDMM)-HMM, the model parameters consist of five major components $\lambda = \{\pi, A, \mu, \Sigma, W\}$, where $\pi = \{\pi_i\}$ denotes the set of initial state probabilities, $A = \{a_{ij}\}$ is the set of state transition probabilities, $\mu = \{\mu_{ik}\}$ is the set of mean vectors of Gaussian mixtures, $\Sigma = \{\Sigma_{ik}\}$ is the set of covariance matrices of Gaussian mixtures , and $W = \{w_{ik}\}$ is the set of intensity weighting of Gaussian mixtures.

The HMM output probability (likelihood) $P(\mathbf{s}|\lambda)$ is defined to be a function of model parameters λ and speech inputs s, i.e., $P(\mathbf{s}|\lambda) = \Psi(\mathbf{s}, \lambda)$. Based on the functional dependencies of the HMM's likelihood to model parameters λ and inputs s, the Baum-Welch inversion of HMM can be derived. More specifically, the Baum-Welch reestimation of an HMM maximizes $P(\mathbf{s}|\lambda)$ by finding the model parameters λ based on a fixed set of speech inputs $\{\mathbf{s}\}$, while the *inversion* of an HMM maximizes $P(\mathbf{s}|\lambda)$ by finding speech inputs $\{\mathbf{s}\}$ that optimize some criterion with given model parameters λ. The Baum-Welch *inversion* is a dual procedure to the Baum-Welch *reestimation* algorithm [4, 5].

In Section 2, the Baum-Welch HMM inversion algorithm is derived and the duality relationship between the Baum-Welch reestimation and inversion is addressed. Section 3 introduces the application of Baum-Welch inversion to robust speech recognition task. Section 4 presents an intensive comparative simulation study of the proposed robust speech classifier under various mismatch conditions.

0-7803-2431-5/95 $4.00

Concluding remarks are given in Section 5.

2. BAUM-WELCH INVERSION OF HIDDEN MARKOV MODEL

2.1. Baum-Welch HMM Inversion

Given the auxiliary function to be maximized:

$$Q(\lambda, \lambda; \mathbf{s}, \mathbf{s}') = \sum_{\theta} \sum_{\mathcal{K}} P(\mathbf{s}, \theta, \mathcal{K}|\lambda) \cdot \log P(\mathbf{s}', \theta, \mathcal{K}|\lambda) \quad (1)$$

where θ and $\mathcal{K}$ denote the possible state transition sequence and the Guassian mixture segmentation sequence, respectively, for a T-frame speech utterance $\mathbf{s} = \{\mathbf{s}_t, 1 \le t \le T\}$ and $\mathbf{s}$ denotes the sequence of old speech features and $\mathbf{s}'$ denotes the sequence of new speech features in speech feature space $\mathcal{S}$.

The problem of the inversion is to find $\bar{\mathbf{s}}$ that maximizes $Q(\lambda, \lambda; \mathbf{s}, \mathbf{s}')$.

$$\bar{\mathbf{s}} = \arg \max_{\mathbf{s}' \in S} Q(\lambda, \lambda; \mathbf{s}, \mathbf{s}')$$

The auxiliary function can be expanded as [4]:

$$Q(\lambda, \lambda; \mathbf{s}, \mathbf{s}') = \sum_{\theta} \sum_{\mathcal{K}} P(\mathbf{s}, \theta, \mathcal{K}|\lambda) \{\log \pi_{\theta_0} + \sum_{t=1}^{T} \log a_{\theta_{t-1}\theta_t}$$

$$+ \sum_{t=1}^{T} \log b_{\theta_t k_t}(\mathbf{s}'_t) + \sum_{t=1}^{T} \log w_{\theta_t k_t}\} \quad (2)$$

where $b_{ik}(\cdot)$ denotes the observation probability for i-th state and k-th mixture.

By equating the derivative of $Q(\cdot)$ with respect to $\mathbf{s}'_t$, i.e., $\frac{\partial Q(\lambda, \lambda; \mathbf{s}, \mathbf{s}')}{\partial \mathbf{s}'_t}$, to be zero,

$$\frac{\partial Q(\lambda, \lambda; \mathbf{s}, \mathbf{s}')}{\partial \mathbf{s}'_t} = \frac{\partial}{\partial \mathbf{s}'_t} [\sum_{\theta} \sum_{\mathcal{K}} P(\mathbf{s}, \theta, \mathcal{K}|\lambda) \sum_{t=1}^{T} \log b_{\theta_t k_t}(\mathbf{s}'_t)]$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} P(\mathbf{s}, i, k|\lambda) \cdot (\mathbf{s}'_t - \mu_{ik}) = 0, \quad (3)$$

we can thus find the reestimated inputs $\bar{\mathbf{s}}_t$.

$$\bar{\mathbf{s}}_t = \frac{\sum_{i=1}^{N} \sum_{k=1}^{K} P(\mathbf{s}, i, k|\lambda) \cdot \mu_{ik}}{\sum_{i=1}^{N} \sum_{k=1}^{K} P(\mathbf{s}, i, k|\lambda)} \quad (4)$$

Note that $P(\mathbf{s}, i, k|\lambda)$ is equivalent to

$$P(\mathbf{s}, i, k|\lambda) = \sum_{j=1}^{N} \alpha_j(t-1) a_{ji} w_{ik} b_{ik}(\mathbf{s}_t) \beta_i(t),$$

where $\alpha_i(\cdot)$ and $\beta_i(\cdot)$ denote the forward and backward probabilities, respectively, therefore the HMM inversion of $\bar{\mathbf{s}}_t$ for an N-state and K-mixture CDMM-HMM is derived

$$\bar{\mathbf{s}}_t = \frac{\sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{j=1}^{N} \alpha_j(t-1) a_{ji} w_{ik} b_{ik}(\mathbf{s}_t) \beta_i(t) \mu_{ik}}{\sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{j=1}^{N} \alpha_j(t-1) a_{ji} w_{ik} b_{ik}(\mathbf{s}_t) \beta_i(t)} \quad (5)$$

2.2. Duality Between HMM Reestimation and HMM Inversion

There is a *duality*, in the sense of maximizing paradigms, between Baum-Welch HMM reestimation and inversion. HMM inversion moves the input speech $\{\mathbf{s}\}$ closer to the mean $\{\mu_{ik}\}$ of a Gaussian mixture by fixing the mean location of each mixture. On the other hand, the HMM reestimation moves the mean $\{\mu_{ik}\}$ location of each Gaussian mixture closer to the input speech $\{\mathbf{s}\}$ by fixing the input speech location. Figure 1 shows the conceptual difference between HMM reestimation and inversion, where mean $\{\mu_{ik}\}$ of each Gaussian mixture is marked as 'o', the noise-free input speech $\{\mathbf{s}\}$ is marked as 'x', and the noisy input speech $\{\mathbf{y}\}$ is marked as '*'. The HMM inversion algorithm moves noisy speech (*) toward mean location (o) of a model. The HMM reestimation algorithm moves mean location (o) toward noisy speech location (*).

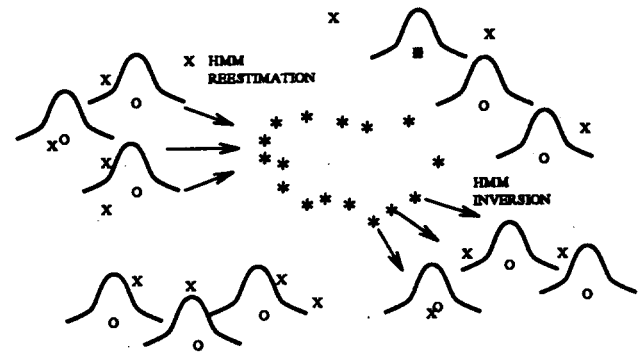

Figure 1: Conceptual difference between Baum-Welch HMM reestimation and HMM inversion.

The HMM inversion quickly converges within a few iterations just as reestimation does. The convergence of HMM inversion is proved theoretically and experimentally in [6].

3. HMM INVERSION FOR NOISY SPEECH RECOGNITION

The HMM inversion proposed in Section 2 can now be applied to classification of noisy speech by adopting the framework of robust hypothesis testing, namely the *robust* HMM inversion. Similar to the MINIMAX technique [1], robust HMM inversion adopts two stage procedure in testing phase. At the first stage, it uses Baum-Welch inversion algorithm to approximate the $\tilde{\mathbf{s}}_m = \arg\max_{\mathbf{s} \in \mathbf{S}_m} P(\mathbf{s}|\lambda_m)$ that maximizes the hypothesized m-th model probability $P(\mathbf{s}|\lambda_m)$, $1 \le m \le \mathrm{M}$, within the mismatch neighborhood $\mathbf{S}_m$. The robustness bound, also utilized by MINIMAX technique, can be effectively employed to avoid *affine phenomenon* [6] which promotes the high similarity of the geometrical shapes between the newly moved noisy speech features and the Gaussian mixture means of any target class. More specifically, the τ-th coefficient of each frame of speech feature, after applying one iteration of the HMM inversion to nominal testing speech $\mathbf{s}_m$, is checked against the inverval $I = \{\mathbf{s}_\tau^{(m)} - R\tau^{-1}\rho^\tau, \mathbf{s}_\tau^{(m)} + R\tau^{-1}\rho^\tau\}$ for some pre-specified constants $R > 0$ and $0 \le \rho < 1$. If the inverted speech $\tilde{\mathbf{s}}_m$ is

within the interval, then $\widetilde{s}_m$ is used as the starting point for next iteration. Otherwise, $\widetilde{s}_m$ is replaced by the end point of I which is closest to inverted speech. At second stage, a model which yield the highest probability $P(\widetilde{s}_m|\lambda_m)$ will be chosen as winner.

Although the robustness bound constraint imposed on HMM inversion technique greatly relaxes the adverse effect caused by affine phenomenon, the extensive bounded movement in HMM inversion can sometimes destroy the original temporal structure of speech rather than moving back to the noise-free speech structure. To further lessen the affine phenomenon, simple noisy speech *scaling* (by a factor of G) procedure is incorporated before the robust HMM inversion so that a *minimal* use of inversion can be assured to maintain the feasible structure of original speech. This scaling before maximization can be likewise applied to the robust MINIMAX classification technique.

To take advantage of both Baum-Welch reestimation and Baum-Welch inversion, the MINIMAX maximization can be combined with the HMM inversion maximization in a batch fashion. More specifically, after completion of MINIMAX maximization which more or less reshapes the original HMM model to be close to the noisy speech, the robust HMM inversion is then performed based on the newly reestimated model $\widetilde{\lambda}$ on the testing speech to fine-tune the speech [6].

Instead of combining the HMM inversion and MINIMAX in a batch fashion, they can be combined in a sequential manner [6]. In this sequential maximization, one iteration consists of a single-step MINIMAX maximization and another single-step HMM inversion. This sequential maximization can be formulated under the framework of Expectation-Maximization (EM) procedure [7], which is an iterative procedure frequently used to solve maximum likelihood (ML) problem in case of *incomplete* data.

4. EXPERIMENTAL RESULTS

Robust HMM inversion is applied to noisy speech recognition to deal with various type of mismatch conditions. The speech database used in this experiment is TI isolated digit database ($M = 10$) which consists of 16 speakers' digit utterances. Ten CDMM-HMMs with $N = 7$ states and $K = 4$ Gaussian mixtures are used to model 10 digits, separately. To get an HMM model which produces the highest recognition performance in noise-free environments, 12-th order LPC cepstral coefficients with frame length of 32 ms and frame shift of 32 ms are used as speech features [1]. The speech samples are pre-emphasized with the filter coefficients 0.97 and hamming windowed before calculating LPC cepstral coefficients. In the training phase, 256 training tokens (16 speakers, 16 repetitions) are used for each HMM digit model. In the testing phase, 300 tokens (6 speakers, 5 repetitions, 10 digits) are used for one experiment and each experiment is repeated 10 times, with different random noise seeds, to get sufficient statistics. Various type of mismatch conditions are simulated, including AWGN, jittering white noise and microphone mismatch. Jittering white noise is generated by multipling the noise standard deviation at each frame with one of five constants [8], i.e., constant={3, 2, 1, 1/2, 1/3}. To simulate the microphone mismatch effect, AWGN corrupted noisy speech data is convolved with a 2nd order FIR filter, $a_1 = -0.45, a_2 = 0.55$.

Table 1 shows the recognition performance of HMMs in noisy environments when the mismatch is incurred by AWGN at various signal-to-noise ratios (SNRs). The performance of HMM which can achieve 95.47% accuracy in noise-free (SNR=∞) environments degrades abruptly to accuracy of 25.73% at SNR of 5 dB without any compensation (see Standard). The robust HMM reestimation with pre-scaling (see Minimax) greatly improves the HMM performance. For example, 36.73% (see Standard) at SNR of 10 dB is increased to 66.37% (see Minimax). It shows consistent performance improvement over the entire SNR. The robust HMM inversion with pre-scaling (see Inversion) also improves the performance of HMM in noisy environments. It outperforms MINIMAX technique at SNR of 15 dB and higher. The slight performance inferiority of the robust HMM inversion vs. robust MINIMAX in low SNR is due to the potential structural distortion caused by inversion process when too much movement is required. The batch combination of robust HMM reestimation and inversion with pre-scaling achieved the highest recognition performance than any other techniques discussed above. For example, 36.73% (see Standard) at SNR of 10 dB is increased to 76.23% (see Batch). Scaling followed by sequential combination (see Sequential) also achieved similar performance to batch combination procedure.

The projection based FOCN (see Projection) showed little improvements due to large distortion of original model of CDMM-HMM. The robustness bound constants (R and ρ) and pre-scaling constant (G) used for this experiment are also shown in Table 1. The constants are roughly fine-tuned empirically.

Table 2 shows the recognition performance of HMMs when testing speech is contaminated by jitter white noise at various level of SNRs. Similar recognition performance for various noisy speech compensation techniques were observed as AWGN contamination.

Table 3 shows the recognition performance of HMMs when different microphone is used for capturing the noisy testing speech at various SNR level. The cepstral shifting compensation is incorporated before the proposed compensation techniques. For the cepstral shifting compensation, the signal-to-noise ratio dependent cepstrum normalization (SDCN) technique [9] is used. After incorporating SDCN technique, the behavior of various compensation technique is similar to the one for AWGN which is explained above in detail (see Table 1). Again, scaling followed by batch or sequential combination of robust HMM reestimation and inversion achieved the best performance.

5. CONCLUSION

The Baum-Welch HMM inversion which exhibits dual property to the Baum-Welch HMM reestimation is applied for robust speech recognition tasks. The robust Baum-Welch inversion and reestimation are found to be very effective in overcoming the affine phenomenon and greatly improve the performance of HMMs under various mismatch conditions. To further reduce the adverse effect of the affine phenomenon, the testing speech is pre-scaled before the ap-

SNR(dB)	5	10	15	20	30	∞
Standard	25.73	36.73	58.90	76.63	91.97	95.47
Minimax	37.73	66.37	81.73	86.80	96.43	96.63
Inversion	35.97	65.40	82.87	88.10	96.33	96.70
Batch	58.23	76.23	84.57	89.20	96.00	96.30
Sequential	57.73	74.93	84.17	89.40	96.30	96.07
Projection	12.73	28.57	50.10	71.23	88.07	91.87
R	4.0	4.0	4.0	2.0	2.0	2.0
ρ	0.3	0.3	0.3	0.3	0.3	0.3
G	1.6	1.6	1.6	1.6	1.6	1.6

Table 1: HMM Performance for White Noise.

SNR(dB)	5	10	15	20	30	∞
Standard	27.43	45.27	62.63	77.63	91.10	95.63
Minimax	43.47	64.60	80.33	85.63	95.73	96.30
Inversion	45.00	65.97	82.27	87.27	95.93	96.83
Batch	60.77	74.90	83.67	89.77	96.23	96.40
Sequential	60.33	75.87	84.87	89.00	95.93	96.33
Projection	19.47	37.33	56.83	73.53	87.60	91.80
R	4.0	4.0	4.0	2.0	2.0	2.0
ρ	0.3	0.3	0.3	0.3	0.3	0.3
G	1.6	1.6	1.6	1.6	1.6	1.6

Table 2: HMM Performance for Jittering Noise.

SNR(dB)	5	10	15	20	30	∞
Standard	19.43	33.63	52.10	71.07	88.90	94.73
Minimax	28.83	55.53	73.27	81.87	93.07	95.33
Inversion	26.97	52.60	72.60	83.47	93.07	95.30
Batch	43.33	63.67	77.13	85.17	92.90	95.20
Sequential	41.80	63.87	77.10	85.67	92.97	95.27
Projection	11.63	18.90	40.77	60.13	83.63	89.93
R	4.0	4.0	4.0	2.0	2.0	2.0
ρ	0.3	0.3	0.3	0.3	0.3	0.3
G	1.6	1.6	1.6	1.6	1.6	1.6

Table 3: HMM Performance for Microphone Mismatch.

plication of robust Baum-Welch inversion (or reestimation) so that best use of inversion (or reestimation) is guaranteed when it is critically needed. And it further increases the performance of HMMs than without scaling.

Combination of robust Baum-Welch reestimation and robust Baum-Welch inversion substantially increased the recognition performance of HMMs under various mismatch conditions. Especially, the sequential combination procedure has rigorous theoretical support from the Expectation-Maximization (EM) procedure.

For future research, different optimization criteria other than maximum likelihood (ML), e.g., maximum mutual information (MMI) [10], and minimum classification error (MCE) [3], based on gradient method [11, 12] can be investigated. Furthermore new constraints, such as a smoothness constraint across time frames of testing speech, can be investigated.

6. REFERENCES

[1] Neri Merhav, Chin-Hui Lee. A minimax classification approach with application to robust speech recognition., IEEE Trans. on SAP, Vol. 1, No. 1, pp. 90-100, January 1993.

[2] B. H. Juang, K. K. Paliwal. Hidden Markov models with first-order equalization for noisy speech recognition., IEEE Trans. on SP, Vol. 40, No. 9, pp. 2136-2143, September 1992.

[3] S. Katagiri, C. H. Lee, B. H. Juang. New discriminative training algorithm based on the generalized probabilistic descent method., Proc. IEEE Workshop Neural Network for Signal Processing, pp. 299-308, Piscataway NJ, August 1991.

[4] B. H. Juang. Maximum-likelihood estimation for mixture multivariate stochastic observations of Markov chains., AT&T Technical Journal., Vol. 64, No. 6, pp. 1235-1249, July 1985.

[5] L. R. Rabiner, B. H. Juang. Fundamentals of speech recognition. Prentice-Hall Inc., 1993.

[6] Seokyong Moon, Jenq-Neng Hwang. Robust speech recognition based on inversion of hidden Markov models., Submitted to IEEE Trans. on SAP for publication, December 1994.

[7] A. P. Dempster, N. M. Laird, D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm., J. Royal Stat. Soc., Series 39, pp. 1-38, 1977.

[8] B. A. Carlson, M. A. Clements. A projection-based likelihood measure for speech recognition in noise., IEEE Trans. on SAP, 2(1):97-102, January 1994.

[9] A. Acero, R. M. Stern. Environmental robustness in automatic speech recognition., IEEE Int'l Conference on ASSP, pp. 849-852, April 1990.

[10] P. F. Brown. Acoustic-phonetic modeling problem in automatic speech recognition., Ph.D. Thesis, Carnegie Mellon University, 1987.

[11] J. N. Hwang, C. H. Chan. Iterative constrained inversion of neural networks and its applications., In Proc. 24-th Conf. on Information Systems and Sciences, pp. 754-759, Princeton, March 1990.

[12] Yoshua Bengio, Renato De Mori, Giovanni Flammia, Ralf Kompe. Global optimization of a neural network-hidden Markov model hybrid., IEEE Trans. on NN, Vol. 3, No. 2, pp. 252-259, March 1992.