JOINT MODEL AND FEATURE SPACE OPTIMIZATION FOR ROBUST SPEECH RECOGNITION

Jenq-Neng Hwang, *Chien-Jen Wang*

Information Processing Laboratory
Department of Electrical Engr., Box # 352500
University of Washington, Seattle, WA 98195, USA
hwang@ee.washington.edu, cjw@pierce.ee.washington.edu

ABSTRACT

This paper presents a maximum likelihood joint-space adaptation technique for robust speech recognition. In the joint-space adaptation process, the *N-Best* hidden Markov model (HMM) inversion frame-by-frame adapts the speech features non-parametrically to compensate the temporal deviation, while the models are transformed parametrically to catch the global characteristics of the mismatch. The proposed joint-space adaptation provides a better compensation to the mismatch than either of the single-space adaptation does. This algorithm operates only on the given testing speech and the models, therefore no adaptation data are required. As verified by the experiments performed under different mismatch environments, the proposed method improves the performance in all the cases without degrading the performance under the match condition.

1. INTRODUCTION

The performance of automatic speech recognizer degrades drastically when the recognizer is deployed under the environments mismatched to the training environment. From the perspective of the temporal evolution, some mismatches cause similar distortion to each frame of the testing speech and can therefore be formulated as a time invariant mapping which can be compensated statically. On the other hand, some mismatches are varying between frames and result in a time-varying mapping, thus the dynamic compensation should be incorporated. Due to the random characteristics of the mismatch, an effective mismatch compensation should be able to both statically and dynamically deal with the features or models.

Most static and dynamic methods operate only in the single space, either the feature-space or the model-space. If the recognizer is tested under severe environments, the highly nonlinear distortion is difficult to compensate with only a single-space modification. This motivates the use of joint-space modifications. In [5], the HMM inversion which provides a frame-by-frame based feature vector modification, is combined with minimax approach [4] to model modification for a joint-space adaptation. Since only the testing speech and models are required therefore this algorithm can recognize testing speech from most unknown environments. This method has been successfully applied to isolated speech recognition tasks. The goal of this paper is to further extend this joint-space adaptation into continuous speech recognition problems.

This paper is organized as follows: Section 2 introduces the extension of HMM inversion procedure for continuous speech recognition. In section 3, the joint feature- and model-space adaptation based on this extension is formulated. The implementation issues are also considered. The experimental results are reported in Section 4. The conclusions are summarized in Section 5.

2. CONTINUOUS SPEECH HMM INVERSION

2.1. EM Algorithm for HMM Inversion

To apply the HMM inversion to the continuous speech recognition problem, a specified word must be assigned to each individual frame of the testing utterance. The inversion process estimates the speech feature vectors $\hat{\mathbf{s}}$ by maximizing the likelihood of the speech and the word sequence given the models. With the EM framework, the HMM inversion for continuous speech is formulated as follow: In the E-step, the auxiliary function is formed as:

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

In the M-step, the speech is estimated as:

$$\hat{\mathbf{s}} = \arg\max_{\mathbf{s} \in \mathcal{S}} Q(\mathbf{s}, \mathbf{s}', W; \lambda), \quad (2)$$

where λ represents only the models occurred in the word sequence, that is generated from the Viterbi search algorithm [2] to provide a meaningful target. To each time index, the state through which the Viterbi path passes dominates the likelihood among the total likelihood of all states. The word sequence can be replaced by the most likely state sequence to serve as the target. The estimated speech $\hat{\mathbf{s}}$ can thus be derived as:

$$\hat{\mathbf{s}} = \arg\max_{\mathbf{s} \in \mathcal{S}} \{\arg\max_{\nu \in W} P(\mathbf{s}, \mathbf{s}', \nu|\lambda)\}. \quad (3)$$

This highly reduces the computational burden. The auxiliary function for the state-dependent inversion can now be restated as:

$$Q(\mathbf{s}, \mathbf{s}', \nu; \lambda) = \sum_{\mathcal{K} \in \nu} P(\mathbf{s}, \mathcal{K}|\lambda) \cdot \log P(\mathbf{s}', \mathcal{K}|\lambda), \quad (4)$$

where ν is the Viterbi state sequence. By equating the derivative of $Q(\mathbf{s}, \mathbf{s}', \nu; \lambda)$ with respect to $\mathbf{s}'_t$ of each time index to be zero, we can find the reestimated input $\hat{s'}_t$:

$$\begin{aligned}\frac{\partial Q(\mathbf{s},\mathbf{s}',\nu;\lambda)}{\partial s'_t} &= \frac{\partial}{\partial s'_t}\sum_{\mathcal{K}\in\nu} P(\mathbf{s},\mathcal{K}|\lambda)\cdot\sum_{t=1}^{T}\log b_{\nu_t k}(s'_t)\\ &= \sum_{k=1}^{K} P(s_t,\nu_t,k|\lambda)\cdot\mathbf{R}^{-1}_{\nu_t k}\cdot(s'_t-\mu_{\nu_t k})\\ &= 0, \qquad (5)\end{aligned}$$

where $b_{\nu_t k}(\cdot)$ denotes the observation probability of the k-th mixture at the ν_t-th state, $\mathbf{R}_{\nu_t k}$ denotes the covariance matrix of the k-th Gaussian mixture in the ν_t-th state. The mixture dependent probability of the state ν_t in each time index t is represented as:

$$P(s_t,\nu_t,k|\lambda) = c_{\nu_t k} b_{\nu_t k}(s_t). \qquad (6)$$

Assuming that the covariance matrix $\mathbf{R}_{ik}$ is diagonal, $\hat{s}_t$ can be solved element-by-element and the τ-th element is obtained as:

$$\hat{s}_t(\tau) = \frac{\sum_{k=1}^{K} c_{\nu_t k} b_{\nu_t k}(s_t)\mu_{\nu_t k}(\tau)/r^2_{\nu_t k}(\tau)}{\sum_{k=1}^{K} c_{\nu_t k} b_{\nu_t k}(s_t)/r^2_{\nu_t k}(\tau)}, \qquad (7)$$

where $r^2_{\nu_t k}(\tau)$ is the τ-th element on the diagonal of the covariance matrix $\mathbf{R}_{\nu_t k}$. Proper constraints must be imposed to confine the movement inside the mismatch neighborhood to avoid the affine phenomenon [5].

2.2. N-Best Search for HMM Inversion

The HMM inversion process moves the feature vectors closer to the weighted average of the mixture means of their corresponding target states, therefore the state sequence generated from the modified speech in the new search will still be the same as the original one, only with higher likelihood without correcting the mistakes made in the initial search. Due to the high confusion created by the mismatched speech, the most likely state sequence derived from the Viterbi search actually contains misrecognized words. Instead of solely relying on the initially searched word sequence, we apply the N-best search algorithm to find the N most likely word sequences [8]. These different sequences provide alternative choices to the confusion parts of the speech. Therefore, it offers N different target sequences for the HMM inversion to compensate the mismatches. The N-best HMM inversion algorithm can be summarized as:

1. Choose the top N likely word sequences: $\mathcal{W} = \{W^{(1)},\cdots,W^{(n)},\cdots,W^{(N)}\}$.
2. Perform the HMM inversion to each word sequence to find the estimated speech, respectively:

   $$\hat{\mathbf{s}}^{(n)} = \arg\max_{\mathbf{s}\in\mathcal{S}} P(\mathbf{s},\mathbf{s}',W^{(n)};\lambda^{(n)})$$

   and record the new joint likelihood, $P(\hat{\mathbf{s}}^{(n)},W^{(n)}|\lambda^{(n)})$, of the n-th sequence after the HMM inversion, where $\lambda^{(n)}$ denotes the models appear in the word sequence $W^{(n)}$.
3. Select the one with the highest likelihood among the hypotheses to be the winning sequence:

   $$W^{*} = \arg\max_{n} P(\hat{\mathbf{s}}^{(n)},W^{(n)}|\lambda^{(n)}).$$

3. JOINT MODEL AND FEATURE SPACE ADAPTATION

The mismatch compensation can also be dealt with in the model-space. In the model-space compensation, we consider to compensate the "global" mismatch statically. Since only the testing speech is available, a simple linear transform of the Gaussian means of the mixtures based on a diagonal matrix $\mathbf{H}$ for scaling and a bias vector $\mathbf{f}$ for shifting, i.e., $\hat{\mu} = \mathbf{H}\cdot\mu+\mathbf{f}$, is considered. The parameters $\{\mathbf{H},\mathbf{f}\}$ can be solved by adopting the EM algorithm [3]. However, under severe testing environment, the mismatch is difficult to be caught by a simple transformation. We further utilize the the N most likely paths to the model adaptation. With each hypothesized word sequence, the sequence dependent transformation is estimated and models are adapted with this transformation. The likelihood of testing speech given this word sequence is re-evaluated based on the modified models. The sequence resulting in the highest likelihood is the winner.

Although the model-space adaptation technique has been reported to adapt the model parameters efficiently [7], it cannot compensate the deviation in the temporal continuity in fine details. On the other hand, HMM inversion provides the ability to adaptively compensate the mismatch speech to the fine details of the continuity. To reach a better compromise, the model transformation is combined with the HMM inversion so that a joint-space adaptation can be achieved [5]. In this joint-space adaptation technique, the auxiliary function $Q(\mathbf{s},\mathbf{s}';\lambda,\lambda')$ is jointly maximized with respect to $\{\mathbf{s},\lambda\}$.

$$Q(\mathbf{s},\mathbf{s}';\lambda,\lambda') = \sum_{\theta}\sum_{\mathcal{K}} P(\mathbf{s},\theta,\mathcal{K}|\lambda)\cdot\log P(\mathbf{s}',\theta,\mathcal{K}|\lambda'), \qquad (8)$$

where $\mathbf{s}$ and $\mathbf{s}'$ denote the sequences of original and modified speech features in the speech feature-space $\mathcal{S}$, and λ and λ' denote the old and new model parameters in the model parameter space Λ. The optimization problem is to find $\{\hat{\mathbf{s}},\hat{\lambda}\}$ that maximizes $Q(\mathbf{s},\mathbf{s}';\lambda,\lambda')$ in the joint-space of $\{\mathcal{S},\Lambda\}$. In practice, the close form solution is difficult to solve. Hence, it is approximated by maximizing $Q(\mathbf{s},\mathbf{s}';\lambda,\lambda')$ with respect to one space at a time, i.e., each space is independently optimized. The independent optimization for each space is then combined in a sequential manner. This joint-space adaptation procedure is summarized as follows:

1. Choose the top N likely word sequences: $\mathcal{W} = \{W^{(1)},\cdots,W^{(n)},\cdots,W^{(N)}\}$.
2. Perform the joint-space adaptation to each candidate word sequence.

   (a) Perform the model adaptation with respect to each word sequence $W^{(n)}$:

   $$\hat{\lambda}^{(n)} = \arg\max_{\lambda^{(n)}\in\mathbf{\Lambda}^{(n)}} P(\lambda^{(n)},\lambda^{(n)'},W^{(n)};\mathbf{s})$$

   where $\lambda^{(n)}$ denotes the models occurred in the word sequence $W^{(n)}$, $\mathbf{\Lambda}^{(n)}$ denotes the mismatch neighborhood of $\lambda^{(n)}$.

(b) Perform the HMM inversion with the modified model $\hat{\lambda}^{(n)}$, to find the estimated speech:

$$\hat{\mathbf{s}}^{(n)} = \arg\max_{\mathbf{s}\in\mathcal{S}} P(\mathbf{s}, \mathbf{s}', W^{(n)}; \hat{\lambda}^{(n)})$$

(c) Go back to step (a) unless some preset requirement is fit.

(d) Record the new joint likelihood for the n-th sequence: $P(\hat{\mathbf{s}}^{(n)}, W^{(n)}|\hat{\lambda}^{(n)})$. Note that the ordering of Step (a) and Step (b) can be reversed.

3. Select the sequence with the highest likelihood among the N hypotheses to be the winning sequence:

$$W^* = \arg\max_n P(\hat{\mathbf{s}}^{(n)}, W^{(n)}|\hat{\lambda}^{(n)}).$$

Some implementation issues are considered following:

Robust Scaling Factor: Although the robust constraint imposed on HMM inversion relaxes the adverse effect caused by the affine phenomenon, the temporal continuity of the testing speech can be destroyed by extensive movement of HMM inversion. To lessen the affine phenomenon, a static compensation by multiplying a constant to every frame of the noisy speech feature vectors before applying the modification is incorporated. This pre-scaling procedure enables the original temporal continuity of the testing speech to be preserved as much as possible after HMM inversion. The scaling factor ϵ is computed as:

$$\epsilon = 1.6 \cdot \frac{\sum_{t=1}^{T} \|\mu_t\|}{\sum_{t=1}^{T} \|s_t\|}, \tag{9}$$

where μ_t is the mean vector of the Gaussian which is closest to the feature vector s_t in that state which the Viterbi path passes through.

Cepstral Mean Subtraction (CMS) The CMS [1] was originally proposed for removing the bias vector caused by the convolutive corruption. However, the distribution of cepstral coefficients is also disturbed by the additive noise. It was observed that the means of the distribution of the cepstral vectors shifts under additive white Gaussian noise environments [6]. By incorporating of the CMS, which is regarded as a static compensation "before" any of our proposed methods, the bias could be globally reset without disturbing the temporal continuity of the testing speech features and improve recognition results.

4. SIMULATION RESULTS

The database used in the experiment is the adult-male part of the TI connected digit corpus. The vocabulary are 10 digits, $0, 1, \cdots, 9$, and o. The lengths of the sequences vary from one to seven, without six. The database is divided into training and testing part by the distributor. All the sequences in the training part, uttered by 55 speakers with total 4235 sequences of all lengths, are used for training. In the testing part, there are 56 speakers. We only choose the 7-digit sequences with total 615 sequences for testing.

The digitized strings were sampled at 20 kHz. In our experiments, the digit sequences are first filtered by a low-pass filter with cutoff frequency 4KHz, then downsampled to 8KHz. The downsampled speech is pre-emphasized with the filter coefficient 0.97 then blocked and Hamming windowed into frames with 32-ms long and 16-ms overlap. For each frame, a 39-dimensional feature vector is extracted: 12-order LPC cepstral coefficients, 12-order delta cepstral coefficients, 12-order delta-delta coefficients, a log-energy coefficient, a delta-log-energy coefficient and a delta-delta-log-energy coefficient. The cepstral coefficients are weighted by the band-pass lifter window with order 12. Each digit is modeled by a single hidden Markov model. The structure of all HMMs are the same: 8-state left-to-right model with non-skip transient probabilities, 5 mixture Gaussians in each state. The long silence at the head and tail of each string is removed. No silence model is incorporated to cope the silence between digits. In our experiment, the string length is assumed to be known. To evaluate the performance, both the string and the word accuracy are computed by **HResults** in HTK1.3 [9].

Various types of corruptions are conducted to the testing speech. These include additive white Gaussian noise (AWGN), additive jittering white noise (AJWN), and simulated microphone mismatch (SMM) [5]. With AJWN corruption, half of the all frames are corrupted by the non-stationary noise. In SMM environment, the degradation of the microphone mismatched speech is different in different frequencies.

There are several methods compared against each others: the no compensation (Standard), the N-best inversion for feature-space adaptation (N-best Inversion), the affine model adaptation for model-space Gaussian-mean adaptation (Model Adaptation), and the iterative joint-space adaptation based on N-best and Model Adaptation (Joint).

Theoretically, all the elements in the feature vectors and the means of Gaussian mixtures need to be adapted during the compensation. In our experiments, only the cepstral and delta-cepstral coefficients are modified and the rest of 15 coefficients are left intact since the cepstral and delta-cepstral coefficients are more discriminative among all coefficients. In our experiments, the constants of the mismatch neighborhood [4], $I = \{\mathbf{s}(\tau) - R\tau^{-1}\rho^{\tau}, \mathbf{s}(\tau) + R\tau^{-1}\rho^{\tau}\}$, of the τ-th cepstral coefficient are chosen to be $R = 1$ and $\rho = 0.3$. The mismatch neighborhood of each coefficient is weighted with the corresponding bandpass window used for bandpass liftering. The radii of the mismatch neighborhood of the delta-cepstral coefficients are chosen to be the same as that of the cepstral coefficients, since the delta-coefficients are derived from the cepstral coefficients.

Table 2 shows the recognition performance when the test speech is corrupted by AWGN at different SNRs. The first column of each SNR shows the string accuracy while the second column shows the digit accuracy. All the results shown below have included robust scaling and CMS pre-adaptation. In the N-best inversion, the best "10" candidate strings derived from the N-best search [8] are used in the experiment. The string accuracy degrades to 54.63% at SNR of 20 dB without any compensation. For the single-space adaptation, the model-space adaptation per-

	String accuracy	Digit accuracy
Standard	95.93	99.30
CMS Alone	95.77	99.30
Joint	95.93	99.33

Table 1. HMM performance of clean test speech. Recognition rate is on percentage.

SNR(dB)	20		30		40	
Standard	54.80	90.34	82.60	96.72	92.85	98.68
CMS Alone	64.55	92.66	86.99	97.28	93.01	98.68
N-best Inversion	70.08	93.45	87.15	97.30	93.01	98.70
Model Adaptation	71.54	93.64	87.80	97.47	93.98	98.79
Joint	77.07	94.60	89.59	97.75	94.31	98.86

Table 2. HMM performance under AWGN environment. Recognition rate is on percentage.

forms better than the feature-space adaptation. It is due to the potential deviation of the temporal continuity caused by HMM inversion process when too much fine-detail movement is required. This is consistent with the observation [7] that model-space adaptation achieves better error reduction than feature-space adaptation does.

In the joint-space adaptation, normally three iterations are enough to compensate the mismatch. In addition, in order to evaluate the influence of these processes under match condition, the clean testing speech are also tested with these adaptation techniques. The results are given in Table 1.

Table 3 shows the recognition performance when the testing speech is corrupted by jittering noise. Due to the highly random behavior of jittering noise, the HMM inversion performs poorer than model-space adaptation in high SNR cases as observed in the AWGN case. In such cases, different mismatch neighborhoods are necessary, this calls for a potential future research topic. However, the joint-space adaptation still shows significant performance improvement. Table 4 shows the recognition performance when microphone mismatch is encountered at various SNR levels. Since the noisy speech is corrupted by the convolutive filter, the incorporation of CMS greatly reduce the mismatch especially in low SNR case. Again, it shows that the joint-space adaptation performs better than single-space adaptation.

5. CONCLUSION

In this paper, the joint-space adaptation technique for robust continuous speech recognition is presented and evaluated. This technique is found to be effective in compensating mismatches under different operation environments and greatly improves the performance of recognition. In our proposed method, only the testing speech is needed to adapt the models, in the meantime itself is adapted to reduce the mismatch. Thus, this technique can be performed without any adaptation or stereo data. It is particularly suitable to operate in the fast changing or unknown field environments. Experiments show that the proposed algorithm improves the performance under different mismatch environments. In the match environment, it performs equally well.

SNR(dB)	20		30		40	
Standard	44.23	87.34	79.02	96.05	89.59	98.14
CMS Alone	54.96	89.94	83.41	96.68	90.08	98.75
N-best Inversion	62.38	91.53	86.67	97.21	92.68	98.63
Model Adaptation	62.44	91.60	86.83	97.14	93.50	98.72
Joint	66.67	92.31	88.46	97.58	93.66	98.79

Table 3. HMM performance under AJWN noise environment. Recognition rate is on percentage.

SNR(dB)	20		30		40	
Standard	46.67	86.64	78.70	95.42	92.20	98.51
CMS Alone	59.02	91.41	85.04	97.00	92.52	98.47
N-best Inversion	64.88	91.82	85.04	96.89	92.85	98.58
Model Adaptation	67.32	92.30	85.36	97.08	92.85	98.61
Joint	69.11	92.68	87.32	97.32	93.98	98.82

Table 4. HMM performance under SMM environment. Recognition rate is on percentage.

REFERENCES

[1] S. Furui, "Cepstral analysis technique for automatic speaker verification", IEEE Trans. on ASSP, Vol.29, No.2, pp.254-272, February 1981.

[2] C.H. Lee and L.R Rabiner,"A frame-synchronous network search algorithm for connected word recognition", IEEE Trans. on ASSP, Vol.37, No.11, pp.1649-1658, November, 1989.

[3] C.J. Leggetter and P.C. 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] N.Merhav, and C.H. Lee, "A minimax classification approach with application to robust speech recognition", IEEE Trans. on SAP, Vol. 1, No. 1, pp. 90-100, January 1993.

[5] S.Y. Moon and J.N. Hwang, "Robust speech recognition based on joint model and feature space optimization of hidden Markov models" IEEE Trans. on NN, March, 1997 (in press).

[6] J.P. Openshaw and J.B. Mason, "The Limitations of Cepstral Features in Noise", IEEE Int'l Conference on ASSP, Vol.II pp.49-52, Adelaide, SA, Australia, April 1994.

[7] A. Sankar, L. Neumeyer, and M. Weintraub "An experimental study of acoustic adaptation algorithms", IEEE Int'l Conference on ASSP, pp.713-716, Atlanta, GA May 1996.

[8] F.K. Soong and E.F. Huang, "A tree-trellis based fast search for finding the N-best sentence hypotheses in continuous speech recognition", IEEE Int'l Conference on ASSP, pp.705-708, Toronto, Canada, 1991.

[9] S. Young, "HTK: Hidden Markov Model Toolkit V1.3", Cambridge University Engineering Department.