A MISMATCH-AWARE STOCHASTIC MATCHING ALGORITHM FOR ROBUST SPEECH RECOGNITION

[1]*Yuan-Fu Liao,* [2]*Jeng-Shien Lin and* [3]*Sin-Horng Chen*

[1,2]Department of Electronic Engineering & Institute of Computer and Communication,
National Taipei University of Technology, 1, Sec. 3, Chung-Hsiao E. Rd. Taipei 106, Taiwan
[3]Department of Communication Engineering, National Chiao Tung University
1001 Ta Hsueh Road, Hsinchu, Taiwan
[1]yfliao@ntut.edu.tw, [3]schen@mail.nctu.edu.tw

ABSTRACT

In this paper, we present a mismatch-aware stochastic matching (MASM) algorithm to alleviate the performance degradation under mismatched training and testing conditions. MASM first computes a reliability measure of applying a set of pre-trained speech models to a mismatch test utterance along the time axis or among different feature vector components. It then estimates and compensates the mismatch using the reliability measure to guide the speech segmentation. Experiments on a serious mismatched condition with training on PSTN-speech database and testing on mobile GSM-speech database showed that MASM outperformed the stochastic match (SM) method, especially, for short utterances.

1. INTRODUCTION

In a real-world application, the mismatch between training and testing conditions often results in significant degradation on the performance of an automatic speech recognition (ASR) system. This mismatch may be due to the variations in speaker's characteristics, speaking style, transducer response, channel effect, background noise, and so on. To reduce the mismatch, some form of compensation may result in an improvement on the recognition performance. A comprehensive review of various compensation techniques can be found in [1].

Usually, those compensation schemes use a set of pre-trained speech models to firstly recognize the test utterance using the mismatch-distorted features, and to then estimate and compensate the mismatch between the models and the test utterance. For examples, the signal bias removal (SBR) [2] and stochastic matching (SM) [3] methods use a pre-trained vector quantization (VQ) codebook and the whole set of hidden Markov models (HMM), respectively, to recognize the test utterance in order to estimate and compensate the mismatch.

A drawback of those methods lies in the case that the pre-trained models may not cover the space of the test utterances when the mismatch is serious, and hence give lousy recognition results in the very beginning to trap the compensation methods in a poor local optimum. This is especially true when the test utterance is short.

However, the mismatch distortion may affect the original utterance unequally along the time axis and among different feature vector components. For example, the first- and second-order time derivations of the MFCC (i.e., Δ-MFCC and Δ^2-MFCC) are shown to be less sensitive to channel and noise interferences [4]. The missing feature theory [5] suggests that we could detect and classify the recognition features into reliable and unreliable subsets in order to get rid of unreliable parts.

Based on these ideas, we attempt, in this study, to improve the conventional mismatch-compensation approach by measuring the reliability of applying a pre-trained speech model to recognize the test utterance. Then the reliability measure is utilized to guide the speech segmentation for estimating and compensating the mismatch. The key concepts are stated in the following:

- Assume the mismatch distortion affects the test utterance unequally along the time axis and among different feature vector components.
- Define a divergence measure to evaluate the reliability of applying a pre-trained model to recognize the test utterance along the time axis and for different feature vector components. The divergence measure is further transformed by a smoothed zero-one *sigmoid* function.
- Use the reliability measure to guide the segmentation of the test utterance by the pre-trained speech models for estimating and compensating the mismatch.

The remainder of this paper is organized as follows. In Section 2, the SM algorithm is reviewed and a divergence-based reliability measure is defined. In Section 3, the proposed mismatch-aware stochastic matching (MASM) method is presented in detail. Experimental results showing the efficacy of the proposed method are discussed in Section 4. Finally we summarize our findings in Section 5.

2. THE SM ALGORITHM

2.1. The SM framework and performance issues

The goal of speech recognition is to find a most likely underlying sequence of events $\mathbf{S} = \{S_1, S_2, ..., S_L\}$ embedded in the sequence of distorted observations $\mathbf{Y} = \{y_1, y_2, ..., y_T\}$, given a set of pre-trained models $\mathbf{\Lambda}_X = \{\lambda_{x_i}\}$, where λ_{x_i} is the model of the *i*-th class. When we consider the compensation in feature space, we may assume that the distortion is invertible.

The estimate of the original features $\hat{\mathbf{X}} = \{x_1, x_2, ..., x_T\}$ can hence be obtained by the transform of the corrupted feature $\mathbf{Y}$ via an inverse mismatch distortion function $F_{\mathbf{v}}(\bullet)$, i.e.,

$$\hat{\mathbf{X}} = F_{\mathbf{v}}(\mathbf{Y}), \tag{1}$$

where $\mathbf{v}$ is the parameter vector of the inverse function. The estimated feature $\hat{\mathbf{X}}$ is expected to match better with the pre-trained model $\mathbf{\Lambda}_X$.

The SM [3] algorithm can be employed to estimate the parameter $\mathbf{v}$. This approach is formulated by maximizing the joint likelihood of $\mathbf{Y}$ and $\mathbf{S}$ given the model $\mathbf{\Lambda}_X$ using the Expectation-Maximization (EM) [6] method, i.e.,

$$\hat{\mathbf{v}} = \arg\max_{\mathbf{v},\mathbf{S}} P(\mathbf{Y} \mid \mathbf{v}, \mathbf{S}, \mathbf{\Lambda}_X) \tag{2}$$

This joint maximization over the variables $\mathbf{v}$ and $\mathbf{S}$ may be done iteratively by keeping $\mathbf{v}$ fixed and maximizing over $\mathbf{S}$ and then keeping $\mathbf{S}$ fixed and maximizing over $\mathbf{v}$ [3].

However, the performance of SM may be seriously affected by the accuracy of the initial condition. When the mismatch between $\mathbf{Y}$ and $\mathbf{\Lambda}_X$ is large, the initial estimate of $\mathbf{S}$ may be totally incorrect to let the iteration be trapped in a local optimum. This motivates our study in this paper to develop a divergence-based reliability measure to provide SM a better initial condition.

2.2. Divergence-based reliability measure

The symmetric divergence (or called Jeffrey's distance) [6] is used to measure the probabilistic distance between the feature distribution of a test utterance and the distribution of a speech model of a recognizer. The divergence of a distribution $\boldsymbol{p}$ with respect to another distribution $\boldsymbol{q}$ is defined as [7]

$$D(p \parallel q) = \int [p(x) - q(x)] \log\left(\frac{p(x)}{q(x)}\right) dx \tag{3}$$

The divergence is greater than or equal to zero, and equals zero when the two distributions are identical. In the case of multivariate Gaussian distribution, the divergence between two distributions, $p = \mathbb{N}(\mathbf{u}_p, \mathbf{\Sigma}_p)$ and $q = \mathbb{N}(\mathbf{u}_q, \mathbf{\Sigma}_q)$, becomes [7]

$$D(p \parallel q) = \frac{1}{2}\begin{Bmatrix} (\mathbf{u}_q - \mathbf{u}_p)^T \left(\mathbf{\Sigma}_p^{-1} + \mathbf{\Sigma}_q^{-1}\right)(\mathbf{u}_q - \mathbf{u}_p) \\ +tr(\mathbf{\Sigma}_p^{-1}\mathbf{\Sigma}_q + \mathbf{\Sigma}_q^{-1}\mathbf{\Sigma}_p - 2 \cdot \mathbf{I}) \end{Bmatrix} \tag{4}$$

To convert the divergence measure into a reliability measure of applying the speech model to various mismatch conditions, we embed the divergence measure into a smoothed zero-one function, e.g., the *sigmoid* function. The reliability measure is then defined as follows:

$$R(D) = \frac{2}{1 + \exp(\alpha D))}, \tag{5}$$

where α is a scaling parameter of the *sigmoid* function. Thus, we expect to measure the reliability of applying a speech model under various mismatch environments by Equation 5 in term of input speech features and the pre-trained speech models.

If the speech recognizer utilizes only diagonal covariance matrices, scalar forms of Equation 4 and 5 can be used to calculate the reliability measures for different feature vector components, i.e.,

$$D_d(p_d \parallel q_d) = \frac{1}{2}\begin{Bmatrix} (u_{q_d} - u_{p_d})^2 \left(\frac{1}{\sigma_{p_d}} + \frac{1}{\sigma_{q_d}}\right) \\ +\left(\frac{\sigma_{q_d}}{\sigma_{p_d}} + \frac{\sigma_{p_d}}{\sigma_{q_d}} - 2\right) \end{Bmatrix}, \tag{6}$$

$$R_d(D_d) = \frac{2}{1 + \exp(\alpha D_d))}$$

for $d = 1, \cdots, M$, where M is the dimension of the feature vector and $u_{p_i}, u_{p_i}, \sigma_{p_i}, \sigma_{q_i}$ are, respectively, the means and variances of the *d*-th feature vector component. Moreover, the reliability can also be measured along the time axis, i.e., we could compute $R_{t,d}(D_{t,d})$, where *t* is the time index.

3. THE PROPOSED MASM ALGORITHM

The proposed MASM algorithm is an iterative scheme which consists of three major steps that can be applied multiple times for further improvement. First, it measures the probabilistic distance between the distribution of the features of a test utterance and the pre-trained speech models. The divergence measure is then converted into the reliability measure. Second, the reliability measure is used to guide the recognizer by emphasizing the reliable feature vector components to obtain a better segmentation of the input utterance. Finally the SM method is adopted to estimate and compensate the mismatch using the segmentation and the pre-trained speech models. The detail algorithm iterative in the variable (l) is summarized below:

The MASM Algorithm

Step 1: Calculate divergence and reliability measures

1) Perform the recognition procedure using $\hat{\mathbf{X}}^{(l)}$ and the pre-trained speech model $\mathbf{\Lambda}_X$ to find a segmentation of the input utterance. Here, l is the iteration index.
2) Construct a set of feature distributions $\mathbf{p}^{(l)} = \{p_1, p_2, ..., p_K\}$ for recognized speech units along the time axis or for different feature vector components.

3) Compute the divergence measures $D_{t,d}^{(l)}$ between the feature distributions of the recognized speech units and the distributions of their corresponding pre-trained speech models $\mathbf{q}^{(l)} = \{q_1, q_2, ..., q_K\}$.
4) Convert the divergence measure into the reliability measure $R_{t,d}^{(l)}\left(D_{t,d}^{(l)}\right)$.

Step 2: Reliability measure-guided segmentation
1) Using the reliability measure to guide the speech segmentation using the following likelihood function, Equation 7. It is worth noting that by using Equation 7, the effect of unreliable parts of input speech features could be alleviated.

$$L\left\{\hat{\mathbf{X}}^{(l)}, \mathbf{v}^{(l)}, \mathbf{S} \mid \mathbf{\Lambda}_X\right\} = p(s_0)\prod_{t=1}^{T} p(s_t \mid s_{t-1}) \sum_{n=1}^{N} c_{S_t,n} \prod_{d=1}^{M} \left\{ \begin{matrix} R_{t,d}^{(l)} \cdot \mathbb{N}(x_{t,d}^{(l)} \mid u_{S_t,n,d}, \sigma_{S_t,n,d}) \\ +(1-R_{t,d}^{(l)}) \end{matrix} \right\} \quad (7)$$

2) Find the $(l+1)$-th speech segmentation $\hat{\mathbf{S}}^{(l+1)}$ by

$$\hat{\mathbf{S}}^{(l+1)} = \arg\max_{\mathbf{S}} L\left\{\mathbf{X}^{(l)}, \mathbf{v}^{(l)}, \mathbf{S} \mid \mathbf{\Lambda}_X\right\} \quad (8)$$

Step 3: SM mismatch estimation and compensation
1) Use $\hat{\mathbf{S}}^{(l+1)}$ and $\mathbf{\Lambda}_X$ to estimate the mismatch $\hat{\mathbf{v}}^{(l+1)}$ between $\hat{\mathbf{X}}^{(l)}$ and $\mathbf{\Lambda}_X$.
2) Use $\hat{\mathbf{v}}^{(l+1)}$ to find the estimate $\hat{\mathbf{X}}^{(l+1)}$ of the original features.

4. EXPERIMENTS

4.1. The databases

To examine the proposed method, two databases were used. One is the MAT4500 telephone speech database which was collected from the landline PSTN telephones by the Mandarin across Taiwan (MAT) project [8]. The speech signals were received with a Dialogic D/21H card and digitally recorded with a SoundBlaster card. A sampling rate of 8 kHz was used. The database was divided into 9 subsets, i.e., subset 1 to 9. Among them, nine-tenth of the database subsets 4 to 9 (referred to as MAT4500-Training) was used for training the speech models; the remaining one-tenth of the database subsets (referred to as MAT4500-Test) was used for testing. There are in total 4,323/480 speakers, 86,544/9,767 utterances, 715,466/56,626 syllables for training/testing, respectively.

Another database is a mobile-phone database (referred to as ATC_Mobile). It was collected by the Advanced Technology Center (ATC) of Computer and Communications Labs. (CCL), Industrial Technology Research Institute (ITRI), Taiwan, through wireless GSM mobile phones. The speech signals were received and digitally recorded using only a Dialogic D/41-ESC card. The sampling rate is 8 kHz. There are 186 speakers; each produced about 30 utterances, resulted in total 5,850 utterances, 35,290 syllables, about 8.4 hours recording.

In this study, MAT4500-Training was used for training the speech models, MAT4500-Test and ATC_Mobile were used to evaluate various compensation methods in both matched (i.e., MAT4500-Training vs. MAT4500-Test) and mismatched (i.e., MAT4500-Training vs. ATC_Mobile) situations. It is worth noting that the mismatch between MAT4500 and ATC_Mobile is serious because the recording transducer (Dialogic vs. SoundBlaster cards), channel (PSTN vs. GSM) and telephone sets (telephone vs. mobile phone) are all different.

4.2. The baseline, SBR and SM schemes

Since we are interested in the training-test mismatch compensation problem due to transducer, phone set or channel mismatch, a series of free-syllable decoding experiments were evaluated. In all experiments, continuous density HMM with left-to-right topologies and Gamma duration models were used. The recognizer was gender-dependent with 100 right-context-dependent (RCD) *initials* and 40 context independent (CI) *finals* for each gender. The numbers of states and mixtures were empirically set to 3 and 5 states, each with maximum 32 mixtures, for *initial* and *final* HMMs, respectively. In addition, one single-state silence model with 64 mixtures was used. A 38-dimensional feature vector including 12 MFCC, 12 Δ-MFCC, 12 Δ^2-MFCC, 1 Δ-log-energy and 1 Δ^2-log-energy was used.

To compensate the channel distortion, the SBR and SM methods were utilized. For SBR, two gender-dependent 32-codeword codebooks were trained from MAT4500-Training. For SM, feature-space bias estimate utilizing the whole HMMs and one-stage search was used. Moreover, for both SBR and SM, the mismatch-compensation procedures were also incorporated in the training phase in order to generate more compact HMMs.

The results of the baseline, SBR and SM schemes under matched and mismatched cases are shown in Table 1. From Table 1, SBR greatly improved the performances from 51.9% to 61.5% and from 26.2% to 40.0% for the matched and mismatched situations, respectively. The SM method, which used the temporal information of the whole utterance and the detail HMMs, improved the performance more significantly to 62.1% and 44.3% for matched and mismatched situations, respectively.

4.3. The MASM scheme

First, to show that mismatch could unequally distorted different input feature vector components, 12 static MFCCs were removed from the original 38-diemnsional feature vectors. When only the left 26-dimensional dynamic feature vectors were used, the performance of the systems (referred to as DELTA in Table 1) was found to be higher than the baseline scheme in the matched condition (51.9% vs. 54.2%) and was close to SBR system in the mismatched condition (36.9% vs. 40.0%). These results showed that static MFCCs were seriously distorted by the mismatch distortion and hurt the recognition performance, while the Δ-MFCC and Δ^2-MFCC were not. In theory, if the mismatch

distortion is a constant function, the means of the dynamic features should be completely unaffected [4].

According to the DELTA results, the dynamic features have the potential to give SM better initial speech segmentation to avoid being trapped in a local optimum. A heuristic method to improve the SM method is therefore suggested. In the 1st SM iteration, the segmentation procedure was modified to use only the dynamic features. After the 1st iteration, all 38 features were used. It is worth noting that this modification is similar to change the initial condition of SM by artificially setting the reliability measure to zero for 12 static MFCCs and to one for all other 26 dynamic features. The results in Figure 1 (referred to as SM_DELTA_1 and SM_DELTA_3) show that this heuristic method maintains the performance under the matched situation while improve the performance (from 44.3% to 45.9% and 47.2% while using 1 and 3 bias terms, respectively) in the mismatched condition.

The performance of the proposed MASM method was then evaluated. In Step 1, three distributions for *initial*, *final* and silence were computed for every different feature vector components. The results listed in Table 1 (referred to as MASM_1 and MASM_3), showed that MASM further increased the performance to 46.2% and 48.1%, while using 1 and 3 bias terms, respectively, in the mismatched condition.

Finally, the relationship between system performance and utterance length under mismatched condition was analyzed. As shown in Figure 1, the shorter the utterance was, the more improvement the proposed method achieved. So, the proposed methods could reliably compensate short utterances under serious mismatch condition. In summary, the results in Table 1 and Figure 1 confirmed the efficiency of the proposed method.

5. CONCLUSIONS

We have proposed an MASM algorithm based on a reliability measure to compensate the mismatch between the training and testing conditions. Experiments on a serious mismatched condition with training on PSTN-speech database and testing on GSM-speech database, have verified that:

- different feature vector components are unequally affected by the mismatch;
- SM may perform better than SBR under the mismatched condition;
- MASM could further improve the performance of SM, especially, for those short utterances under the mismatched condition.

One area of further research is to study the reliability measure under various situations, especially, to extend to measure the reliability under noisy condition.

6. ACKNOWLEDGEMENT

This work was supported in part by National Science Council (NSC), Taiwan, under contract NSC-91-2219-E-027-004 and in part by Ministry of Education (MOE) under contract EX-91-E-FA06-4-4. The authors also want to thank The Association for Computational Linguistics and Chinese Language Processing (ROCLING) and ATC of CCL, ITRI, Taiwan for supporting the MAT4500 and ATC_Mobile databases, respectively.

7. REFERENCES

[1] Chin-Hui Lee and Qiang Huo "On adaptive decision rules and decision parameter adaptation for automatic speech recognition," *Proceedings of the IEEE*, vol.88, no.8, pp.1241-1269 August 2000.

[2] Mazin G. Rahim and Bing-Hwang Juang, "Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition, " *IEEE Trans. on Speech and Audio Processing*, vol.4, pp.19-30, January 1996.

[3] Ananth Sankar and Chin-Hui Lee, "A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, vol.4, no. 3, pp.190-202, May 1996.

[4] W.-J. Wang and S.-H. Chen, "Signal bias removal with orthogonal transform for adverse Mandarin Speech recognition," *Electronics Letters*, vol. 36, no. 9, pp. 852-852, April, 2000.

[5] J. Barker, M. Cooke, L. Josifovski and P. Green, "Soft Decisions in Missing Data techniques for Robust Automatic Speech Recognition," *ICSLP 2000*, Beijing.

[6] P. A. Devijver and J. Kittler, "Pattern Recognition – A Statistical Approach," Prentice-Hall International, London, 1982.

[7] A. Dempster, N. Laird and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. Royal Statist. Soc., vol. 39, pp. 1-38, 1977.

[8] H. C. Wang, "MAT - A project to collect Mandarin speech data through telephone networks in Taiwan," *Computational Linguistics and Chinese Language Processing*, vol. 2, no. 1, pp. 73-89, 1997.

Table 1: Experimental results of various methods under matched and mismatched situations.

	Match	MisMatch	# of Biases
MASM_3	-	**48.1%**	3
SM_DELTA_3	-	**47.2%**	3
MASM_1	-	**46.2%**	1
SM_DELTA_1	**62.1%**	**45.9%**	1
SM	62.1%	44.3%	1
SBR	61.5%	40.0%	1
DELTA	54.2%	36.9%	-
BASELINE	51.9%	26.2%	-

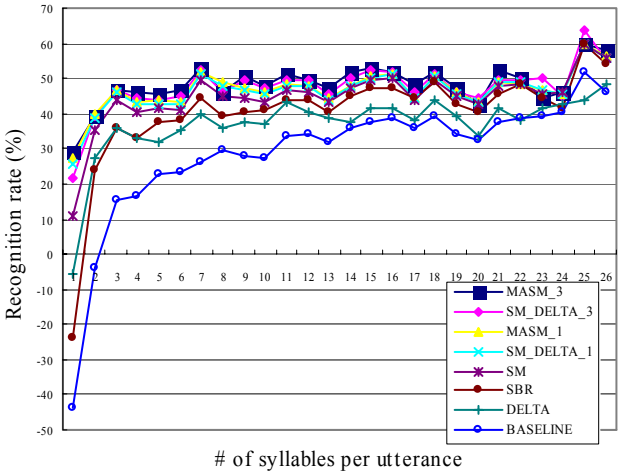

Figure 1: Analysis of the ATC_Mobile recognition results.