ARTIFICIAL STEREO DATA GENERATION FOR SPEECH FEATURE MAPPING

Chang Woo Han, Tae Gyoon Kang, Shin Jae Kang, June Sig Sung, and Nam Soo Kim

School of Electrical Engineering and INMC
Seoul National University, Seoul 151-742, Korea
E-mail: {cwhan, tgkang, sjkang, jssung}@hi.snu.ac.kr, nkim@snu.ac.kr

ABSTRACT

Feature mapping technique is widely used to eliminate the mismatch between the training and test conditions of speech recognition. In the feature mapping, a target (mismatched) feature vector sequence is mapped closer to the corresponding reference (matched) feature vector stream. The training of the mapping system is usually carried out based on a set of stereo data which consists of simultaneous recordings obtained in both the reference and target conditions. In this paper, we propose a novel approach to blind parameter estimation which does not require the reference feature vectors. The proposed approach is motivated by the hidden Markov model (HMM)-based speech synthesis algorithm.

Index Terms— Robust speech recognition, feature mapping, blind estimation

1. INTRODUCTION

In general, the performance of a speech recognition system degrades when there is a mismatch between test and training conditions. There are several factors that lead to acoustic mismatch such as the background noise, different audio devices, reverberations, data compression modules, etc. In order to ameliorate the degradation in recognition performance, feature mapping techniques have been frequently applied [1]-[9]. In the feature mapping techniques, the signal waveforms or feature vectors are enhanced during front-end processing.

Depending on the type of training or adaptation data, parameter estimation approaches for feature mapping can be divided into stereo-based and blind techniques. Stereo-based technique is applied when there exists a database of simultaneous recordings obtained in both the reference and target conditions, and feature mapping rules are derived from the difference between the pair of feature vectors [1]-[5]. In the blind technique, on the other hand, only the input feature vectors are given and the information related to the target feature vectors is usually provided by statistical models such as the Gaussian mixture model (GMM), hidden Markov model (HMM) and switching linear dynamic model (SLDM) [6]-[8]. In general, feature mapping for the blind technique is done based on either the minimum mean square error (MMSE) or the maximum likelihood (ML) criterion. In our recent study, we proposed a stereo-based feature mapping approach based on the switching linear dynamic system (SLDS) [4], [5]. One of the prominent advantages of the SLDS is that it enables a systematic implementation of sequence-to-sequence mapping instead of the traditional vecto r-to-vector mapping [3].

In this paper, we propose an approach to blind estimation for the speech feature mapping algorithms which originally require stereo data for their parameter training. In the proposed method, an artificial reference feature vector sequence are generated from the HMM and then applies it to a conventional stereo-based technique. Our approach is motivated by the speech feature generation method employed in HMM-based speech synthesis [10]. In order to further improve the performance of the feature mapping system, we also propose to interpolate the feature vector streams generated through the HMM with those obtained from the output of a conventional feature compensation algorithm. The proposed blind estimation technique was applied to a task of speech recognition over the Aurora-5 DB and has demonstrated a remarkable performance improvement.

2. STEREO-BASED FEATURE MAPPING

Suppose that we have two simultaneous recordings of the same speech realizing a word sequence: one is obtained in the target (mismatched) and the other in the reference (matched) conditions. Let $\mathbf{x}_1^T = (\mathbf{x}_1' \ \mathbf{x}_2' \ \cdots \ \mathbf{x}_T')'$ be the sequence of feature vectors of length T extracted from the recording obtained in the target condition with the prime denoting the transpose of a vector or a matrix, and $\mathbf{x}_t \in R^d$ represent the feature vector at time t. In a similar way, $\mathbf{y}_1^T = (\mathbf{y}_1' \ \mathbf{y}_2' \ \cdots \ \mathbf{y}_T')'$ represents the corresponding sequence of feature vectors obtained in the reference condition. In the feature mapping approaches, a feature vector sequence $\mathbf{x}_1^T$ obtained in the mismatched condition is mapped to a feature sequence $\hat{\mathbf{y}}_1^T = (\hat{\mathbf{y}}_1', \hat{\mathbf{y}}_2', \cdots, \hat{\mathbf{y}}_T')'$ which is considered a promising counterpart in the matched condition.

This research was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 2011-0020407).

A variety of feature mapping techniques have been proposed in the past to compensate the mismatch between the training and test conditions. Recently, we proposed the SLDS-based feature mapping technique, which systematically implements a sequence-to-sequence mapping in contrast to the conventional vector-to-vector mapping approaches [4]. In this section, we briefly review the SLDS which is a sequence-to-sequence mapping technique including most of the conventional vector-to-vector mapping approaches as its special cases [5].

In the SLDS, the output feature vector sequence $\mathbf{y}_1^T$ is assumed to be generated from the input feature vector stream $\mathbf{x}_1^T$ by switching K different linear dynamic systems (LDS's) [5]. When the k-th LDS is applied, the feature mapping process is approximated by following

$$\mathbf{z}_{t+1} = A_k \mathbf{z}_t + B_k \mathbf{x}_t + \mathbf{m}_{u,k} \tag{1}$$

$$\hat{\mathbf{y}}_t = C_k \mathbf{z}_t + D_k \mathbf{x}_t + \mathbf{m}_{w,k} \tag{2}$$

where $\mathbf{z}_t$ denotes the hidden state of the system at time t and $\lambda_k = \{A_k, B_k, C_k, D_k, \mathbf{m}_{u,k}, \mathbf{m}_{w,k}\}$ are the LDS parameters to be estimated. If the a posteriori probability of each LDS is available, we can employ a soft-decision scheme which modifies (1) and (2) into

$$\mathbf{z}_{t+1} = \sum_{k=1}^{K} p\left(k|\mathbf{x}_t\right) \left[A_k \mathbf{z}_t + B_k \mathbf{x}_t + \mathbf{m}_{u,k}\right] \tag{3}$$

$$\hat{\mathbf{y}}_t = \sum_{k=1}^{K} p\left(k|\mathbf{x}_t\right) \left[C_k \mathbf{z}_t + D_k \mathbf{x}_t + \mathbf{m}_{w,k}\right] \tag{4}$$

where $p\left(k|\mathbf{x}_t\right)$ represents the posterior probability of the k-th LDS. Interested readers are referred to [4], [5] for more detail.

3. ARTIFICIAL STEREO DATA GENERATION

In the stereo-based approaches such as SLDS, in order to estimate the relevant parameters, a set of stereo data has to be given. This means that for each target feature vector sequence $\mathbf{x}_1^T$ we have the corresponding reference feature vector sequence $\mathbf{y}_1^T$. The two feature vector sequences, $\mathbf{x}_1^T$ and $\mathbf{y}_1^T$ are extracted from simultaneous recordings of the same speech. However, in the blind technique, the actual reference feature vector sequence $\mathbf{y}_1^T$ is unavailable and all that we have are the target feature vector sequence $\mathbf{x}_1^T$ and a statistical model for $\mathbf{y}_1^T$. In this section, we propose novel approaches to generate artificial reference feature vector stream. Once the artificial reference feature vector sequence is generated for each target feature vector sequence, a conventional stereo-based technique can be straightforwardly applied to estimate the mapping parameters.

3.1. Artificial reference feature generation from HMM

Suppose that the statistical model for $\mathbf{y}_1^T$ is given by an HMM. Then the HMM, Λ_y which characterizes the statistical properties of $\mathbf{y}_1^T$ is assumed to consist of S states and the observation distribution at each state is given by a Gaussian mixture model (GMM). Conventionally in speech recognition, the HMM Λ_y is defined over an extended feature vector to account for both the static and dynamic characteristics simultaneously. Let $\mathbf{y}_t$ be an original reference static feature vector at time t. Then, the extended feature vector $\tilde{\mathbf{y}}_t$ is formed by appending dynamic features e.g., Δ- and $\Delta\Delta$-cepstra to $\mathbf{y}_t$ as follows:

$$\tilde{\mathbf{y}}_1^T = \begin{bmatrix} \tilde{\mathbf{y}}_1 \\ \tilde{\mathbf{y}}_2 \\ \vdots \\ \tilde{\mathbf{y}}_T \end{bmatrix} = \mathbf{W}\mathbf{y}_1^T = \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \\ \vdots \\ \mathbf{W}_T \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_T \end{bmatrix} \tag{5}$$

where $\mathbf{W}$ is a constant matrix, and

$$\tilde{\mathbf{y}}_t = \mathbf{W}_t \mathbf{y}_1^T \ . \tag{6}$$

Generation of an artificial reference feature vector sequence is motivated by the speech feature generation technique in HMM-based speech synthesis [10]. In HMM-based speech synthesis, the goal is to find an optimal feature vector sequence given the HMM parameters in the ML sense, i.e.,

$$\hat{\mathbf{y}}_1^T = \arg\max_{\mathbf{y}_1^T} \log p\left(\mathbf{y}_1^T|\Lambda_y\right) \ . \tag{7}$$

For a specific state sequence $s_1^T = (s_1, s_2, \cdots, s_T)$ and a mixture component sequence $m_1^T = (m_1, m_2, \cdots, m_T)$, the log likelihood can be calculated due to the relation between $\mathbf{y}_1^T$ and $\tilde{\mathbf{y}}_1^T$ as given by (6) as follows:

$$\begin{aligned} &\log p\left(\mathbf{y}_1^T|s_1^T, m_1^T, \Lambda_y\right) \\ &= -\frac{1}{2}\sum_{t=1}^{T} \left(\mathbf{W}_t\mathbf{y}_1^T - \tilde{\mu}_{s_t,m_t}\right)' \tilde{\Sigma}_{s_t,m_t}^{-1} \left(\mathbf{W}_t\mathbf{y}_1^T - \tilde{\mu}_{s_t,m_t}\right) \\ &\quad + \text{Const.} \end{aligned} \tag{8}$$

where $\tilde{\mu}_{s_t,m_t}$ and $\tilde{\Sigma}_{s_t,m_t}$ indicate respectively mean vector and covariance matrix of m_t-th Gaussian mixture at state s_t. Since it is practically difficult to solve (7) directly, we apply the EM algorithm which iteratively updates the estimate for $\mathbf{y}_1^T$. Let $\bar{\mathbf{y}}_1^T = (\bar{\mathbf{y}}_1' \ \bar{\mathbf{y}}_2' \cdots \ \bar{\mathbf{y}}_T')'$ be the estimate for $\mathbf{y}_1^T$ obtained at the previous iteration. Then, at each iteration of the EM algorithm it is updated in the following way:

$$\hat{\mathbf{y}}_1^T = \arg\max_{\mathbf{y}_1^T} E\left[\log p\left(\mathbf{y}_1^T|s_1^T, m_1^T, \Lambda_y\right)|\bar{\mathbf{y}}_1^T, \Lambda_y\right] \tag{9}$$

where $\hat{\mathbf{y}}_1^T = (\hat{\mathbf{y}}_1' \ \hat{\mathbf{y}}_2' \cdots \ \hat{\mathbf{y}}_T')'$ indicates the updated sequence of the reference feature vectors and $E\left[\cdot\right]$ represents the expectation operation.

In order to solve (9), we first compute the a posteriori probability of each Gaussian component, $\{\gamma_t\left(s, m\right)\}$. It can be efficiently obtained by means of the forward-backward algorithm or can be approximated with the use of the Viterbi

algorithm. After $\{\gamma_t(s,m)\}$ are computed, the updated reference feature vector sequence is derived as follows [10]:

$$\hat{\mathbf{y}}_1^T = \left(\sum_{t=1}^{T} \sum_{s=1}^{S} \sum_{m=1}^{M} \gamma_t(s,m) \mathbf{W}_t' \tilde{\Sigma}_{s,m}^{-1} \mathbf{W}_t \right)^{-1} \times \left(\sum_{t=1}^{T} \sum_{s=1}^{S} \sum_{m=1}^{M} \gamma_t(s,m) \mathbf{W}_t' \tilde{\Sigma}_{s,m}^{-1} \tilde{\mu}_{s,m} \right) \quad (10)$$

where M and S indicate the total number of Gaussians and states in Λ_y, respectively.

3.2. Combination with feature compensation technique

One of the drawbacks of the approach proposed in (10) is that the generated feature vector sequences will tend to become similar if we obtain similar alignments for the HMM states and mixture components even though they show quite different characteristics in the original feature domain. This phenomenon may mislead parameter estimation of the feature mapping techniques.

In order to alleviate this problem, it is useful to apply a feature compensation algorithm where an estimate for the clean speech feature is derived by taking advantage of a speech corruption model. Let $\hat{\mathbf{y}}_t^{\text{FC}}$ denote an estimate for $\mathbf{y}_t$ obtained from a feature compensation algorithm and $\hat{\mathbf{y}}_t^{\text{HMM}}$ be the corresponding vector derived from the HMM as shown in (10). Then, one of the simplest ways to generate the artificial reference feature vector $\hat{\mathbf{y}}_t$ is to interpolate between $\hat{\mathbf{y}}_t^{\text{FC}}$ and $\hat{\mathbf{y}}_t^{\text{HMM}}$ such that

$$\hat{\mathbf{y}}_t = \rho \hat{\mathbf{y}}_t^{\text{FC}} + (1-\rho)\hat{\mathbf{y}}_t^{\text{HMM}} \quad (11)$$

where $\rho \in [0,\ 1]$ is an interpolation weight. It is important that the interpolation weight ρ should account for the variance of $\hat{\mathbf{y}}_t^{\text{FC}}$, which can be treated as a measure of uncertainty for the output of the feature compensation algorithm. Similar strategies are often employed in the uncertainty decoding techniques where the back-end recognition parameters are modified depending on the uncertainty measure provided by the front-end module [11].

4. EXPERIMENTS

Proposed approach was applied to the task of speech recognition with the Aurora-5 DB which was developed to investigate the influence on the performance of automatic speech recognition for a hands-free speech input in noisy room environments [12]. Furthermore in Aurora-5, two test conditions are included to study the influence of transmitting the speech in a mobile communication system. The number of test utterances was 8700 for each test condition.

In the experiments, we focused on the performance of the speech recognition system in a clean training condition. Baseline recognition systems were built based on the clean speech data provided by the G. 712 filtered and non-filtered data sets. The number of utterances used for HMM training was 8623 per data set. In our implementation, we employed the conventional frontend (FE) feature specified in the ETSI standard [13] as the basic feature vectors. A 13-dimensional cepstrum and the corresponding Δ- and $\Delta\Delta$-cepstra were extracted from each frame and used as the feature vector for speech recognition. The word accuracies of the baseline systems are shown in Table 1 for the G. 712 filtered and non-filtered data sets.

We evaluated the performance of the SLDS algorithm [5] with various artificial reference feature vector streams. For the non-filtered data set of Aurora-5 DB, 575 utterances were applied to estimate the SLDS parameters for each separate test condition while 431 utterances were used in the case of G. 712 filtered data set. The number of LDS's was set $K = 128$ and the dimension of the state $\mathbf{z}_t$ in (1) was fixed at 39 which was three times of the cepstrum dimension.

For artificial feature generation from HMM, we applied (10). In the case of feature compensation, we applied the interacting multiple model (IMM) algorithm proposed in [6]. For convenience, we denote the SLDS algorithm with artificial reference feature vector stream generated from HMM by *HMM*, and from IMM by *IMM*. We combined the feature vector streams generated through HMM with those obtained from IMM, which we denote by *HMM+IMM*. The interpolation weight ρ in (11) was set to 0.5 which showed a good performance in our experiments. It is noted that *HMM*, *IMM* and *HMM+IMM* are blind approaches while the conventional SLDS algorithm (denoted by *Stereo-based*) is stereo-based technique. The performance of each algorithm was compared in terms of relative error rate reduction (RERR).

Tables 2 and 3 show the RERR's in each separate environmental and SNR condition, respectively. These results clearly demonstrate that the interpolation between the two sets of feature vectors, one derived from a feature compensation algorithm and the other from HMM, is very useful in generating more realistic artificial reference features. One may consider the results obtained from *Stereo-based* as a performance upper bound for any blind estimation techniques. It is noted that the performance of *HMM+IMM* is almost similar to that obtained from stereo-based parameter estimation.

5. CONCLUSIONS

In this paper, we have proposed a novel approach to blind parameter estimation for speech feature mapping. The proposed approach first generates an artificial reference feature vector sequence from the HMM and interpolates it with the output feature vector stream obtained from a feature compensation algorithm. This interpolation enables not only to faithfully reconstruct the clean speech feature but also to increase the likelihood of the HMM used for speech recognition. Future study will include an optimal combining technique based on

Table 1. Word accuracies (%) of the baseline system for non-filtered and G. 712 filtered test data sets

	Non-Filtered			G. 712 Filtered			
Noise	Interior Noise			Car Noise			Street Noise
SNR (dB)		HFO	HFL		HFC	HFC-GSM	GSM
Clean	99.32	93.30	83.24	99.31	97.41	92.45	97.70
15	81.66	71.46	55.49	90.44	71.96	61.20	81.64
10	56.44	43.97	30.72	70.27	42.92	36.56	58.61
5	27.67	18.14	12.56	41.48	19.51	18.39	27.09
0	11.14	6.42	5.74	20.80	11.41	8.68	3.63

HFO: hands-free in office, HFL: hands-free in living room, HFC: hands-free in car, HFC-GSM: HFC & GSM

Table 2. RERR's (%) for different environments.

	Stereo-based	*HMM*	*IMM*	*HMM+IMM*
Interior	67.07	45.53	68.81	68.72
HFO	51.52	46.03	45.55	55.76
HFL	36.95	23.17	35.64	45.68
Car	77.35	59.01	74.50	76.58
HFC	75.22	46.71	52.38	69.57
HFC-GSM	72.28	53.95	40.10	64.31
Street	54.71	24.60	41.72	51.47

Table 3. RERR's (%) for different SNR's.

	Stereo-based	*HMM*	*IMM*	*HMM+IMM*
Clean	50.07	5.94	2.39	44.97
15 dB	73.10	51.16	67.48	74.61
10 dB	74.64	62.43	70.77	76.36
5 dB	64.15	51.69	60.92	64.62
0 dB	42.49	29.11	34.85	40.16
Average	61.55	42.13	50.00	61.06

the Bayesian framework.

6. REFERENCES

[1] L. Deng, A. Acero, L. Jiang, J. Droppo, and X. Huang, "High-performance robust speech recognition using stereo training data," in *Proc. ICASSP*, Salt Lake City, Utah, pp. 301-304, 2001.

[2] J. Droppo, L. Deng, and A. Acero, "Evaluation of the SPLICE algorithm on the AURORA2 database," in *Proc. Eurospeech*, Aalborg, Denmark, pp. 217-220, 2001.

[3] M. Afify, X. Cui, and Y. Gao, "Stereo-based stochastic mapping for robust speech recognition," *IEEE Trans. Audio, Speech and Language Process.*, vol. 17, no. 7, pp. 1325-1334, Sep. 2009.

[4] C. W. Han, T. G. Kang, D. H. Hong, N. S. Kim, K. Eom, and J. Lee, "Switching linear dynamic transducer for stereo data based speech feature mapping," in *Proc. ICASSP*, Prague, Czech Rep., pp. 4776-4779, May 2011.

[5] N. S. Kim et al., "Speech feature mapping based on switching linear dynamic system," *IEEE Trans. Audio, Speech and Language Process.*, (to appear).

[6] N. S. Kim, "IMM-based estimation for slowly evolving environments," *IEEE Signal Process. Lett.*, vol. 5, no. 6, pp. 146-149, June 1998.

[7] B. Mesot, and D. Barber, "Switching linear dynamical systems for noise robust speech recognition," *IEEE Trans. Audio, Speech and Language Process.*, vol. 15, no. 6, pp. 1850-1858, Aug. 2007.

[8] M. Wölfel, "Enhanced speech features by single-channel joint estimation of noise and reverberation," *IEEE Trans. Audio, Speech and Language Process.*, vol. 17, no. 2, pp. 312-323, Feb. 2009.

[9] N. S. Kim et al., "Blind estimation of feature mapping paremeters," *IEEE Trans. Audio, Speech and Language Process.*, (submitted in Sep. 2011).

[10] K. Tokuda et al., "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, vol. 3, pp. 1315-1318, June 2000.

[11] D. Kolossa, and R. Haeb-Umbach, *Robust Speech Recognition of Uncertain or Missing Data*. Springer-Verlag, 2011.

[12] H. G. Hirsch, "AURORA-5 experimental framework for the performance evaluation of speech recognition in case of a hands-free speech input in noisy environments," Niederrhein Univ. of Applied Sciences, Nov. 2007.

[13] ETSI Std. Document, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; front-end feature extraction algorithm; compression algorithm," ETSI ES 201108 V1.1.3, Sep. 2003.