ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

FACTORED MLLR ADAPTATION FOR HMM-BASED EXPRESSIVE SPEECH SYNTHESIS

June Sig Sung, Doo Hwa Hong, Chul Min Lee, and Nam Soo Kim

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

ABSTRACT

One of the most popular approaches to parameter adaptation in hidden Markov model (HMM) based systems is the maximum likelihood linear regression (MLLR) technique. In our previous work, we proposed factored MLLR (FMLLR) where MLLR parameter is defined as a function of a control parameter vector. We presented a method to train the FMLLR parameters based on a general framework of the expectation-maximization (EM) algorithm. To show the effectiveness, we applied the FMLLR to adapt the spectral envelope feature of the reading-style speech to those of the singing voice. In this paper, we apply the FMLLR to HMM-based expressive speech synthesis task and compare its performance with conventional approaches. In a series of experimental result, the FMLLR shows better performance than conventional methods.

Index Terms— MLLR, MRHSMM, Factored MLLR, expressive speech synthesis, HMM-based speech synthesis

1. INTRODUCTION

It is difficult to create a huge speech database when considering the time and cost required. To overcome the deficiency usually observed in a small-sized speech database, the maximum likelihood linear regression (MLLR) method is proposed [1]. In the MLLR approach, original parameters of the HMM's are mapped to their adapted values via a set of affine transformations estimated from a small amount of data which is called the adaptation data.

Since its invention, the MLLR approach has been enhanced and its applications have been extended through a number of studies [2][3]. In [4]-[6], MLLR is applied to multiple regression hidden semi-Markov model (MRHSMM) to take expressive speech style into consideration. By incorporating both MRHSMM and MLLR, the speech synthesizer can express various emotional styles and levels using only a single matrix defined by the style vector.

In our previous work, we extended the conventional MLLR to the factored MLLR (FMLLR) framework where each MLLR parameter is defined as a function of the control parameter vector [7]. More specifically, each element of the MLLR parameters is given as an inner product between a regression vector and a transformed control vector. To show the effectiveness, we applied the FMLLR approach to adapt the spectral envelope features of the reading-style speech to those of the singing voice and compared its performance with the traditional MLLR approaches.

In this paper, we apply the FMLLR approach to HMM-based expressive speech synthesis. Its performance is evaluated in a series of experiments on expressive speech synthesis, and compared with the conventional MLLR and MRHSMM.

2. MLLR

In conventional MLLR adaptation, the mean vector $\boldsymbol{\mu}_s$ of a particular distribution s of the HMM is transformed to $\widehat{\boldsymbol{\mu}}_s$ by

$$\widehat{\boldsymbol{\mu}}_s = \mathbf{M}\,\boldsymbol{\mu}_s + \mathbf{b} \tag{1}$$

where $\mathbf{M}$ is a $p \times p$ regression matrix and $\mathbf{b}$ is a bias vector. The PDF of the distribution s is assumed to be Gaussian with a mean vector $\boldsymbol{\mu}_s$ and covariance matrix $\sum_s$. For convenience, we further assume that both $\mathbf{M}$ and $\sum_s$ are diagonal. With this assumption, each element of $\boldsymbol{\mu}_s$ can be separately adapted.

The parameters $\mathbf{M}$ and $\mathbf{b}$ are estimated according to the maximum likelihood (ML) criterion, and the expectation-maximization (EM) algorithm is applied to increase the likelihood iteratively. Let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T)$ be the given adaptation vectors that are used to adapt the mean vectors of the HMM. At the E-step, we compute a posteriori probability of the distribution s at each time defined by

$$\gamma_t(s) = Pr(\theta(t) = s|\mathbf{X}, \lambda) \tag{2}$$

where $\theta(t)$ indicates the distribution index at time of t and λ means the current parameters to be adapted. Then, at the M-step, we update the $\mathbf{M}$ and $\mathbf{b}$ in order to maximize the

This research was supported in part by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No. 20110020407) and by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency) (NIPA-2011-C1090-1121-0007).

expectation of the complete data log-likelihood which is given as follows:

$$\left\{\widehat{\mathbf{M}},\ \widehat{\mathbf{b}}\right\} = \arg\max_{\{\mathbf{M},\mathbf{b}\}} -\frac{1}{2}\sum_{t=1}^{T}\gamma_t\left(s\right) \times \left(\mathbf{x}_t - \widehat{\boldsymbol{\mu}}_s\right)^{'} \Sigma_s^{-1}\left(\mathbf{x}_t - \widehat{\boldsymbol{\mu}}_s\right) \quad (3)$$

where the prime denotes the transpose of a vector or a matrix. The solution to (3) is computed by differentiation with respect to each $\mathbf{M}(i)$ and $\mathbf{b}(i)$ which denote i-th row vector of $\mathbf{M}$ and $\mathbf{b}$, respectively. For more details, the readers are referred to [1].

3. MRHSMM

We briefly introduce the MRHSMM technique in this section. In the MRHSMM-based style control technique, each HSMM mean vector is described by performing multiple regression on a low dimensional style vector where each component represents the degree or intensity of a specific style. A base model for synthesis in the MRHSMM technique is an HSMM which incorporates a explicit state duration modeling into HMM. In this paper, we only focus on the adaptation of spectral envelope parameters, hence derivations for pitch and duration modeling would be disregarded.

In MRHSMM, it is further assumed that $\boldsymbol{\mu}_s$ is modeled using multiple regression as

$$\boldsymbol{\mu}_s = \mathbf{H}_s\,\boldsymbol{\xi} \quad (4)$$

where $\boldsymbol{\xi}$ is a control vector of dimension L representing the degree of each style of the speech, and $\mathbf{H}_s$ is a multiple regression matrix of dimension $p \times L$ for the distribution s.

When the training data and corresponding control vectors are given, $\mathbf{H}_s$ can be estimated by using the least squares criterion [5]. In this paper we will address the problem of estimating $\mathbf{H}_s$ under the ML framework. Suppose that $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T)$ are the given data vectors accompanied with their control vectors. Covariance matrices, $\{\Sigma_s\}$ are also assumed to be diagonal. By substituting (4) into (3) with the covariance assumption, $\mathbf{H}_s$ is obtained according to

$$\left\{\widehat{\mathbf{H}}_s\right\} = \arg\max_{\mathbf{H}_s} -\frac{1}{2}\sum_{t=1}^{T}\gamma_t\left(s\right) \times \left(\sum_{i=1}^{p}\frac{\left(x_{t,i} - \mathbf{H}_s(i)\,\boldsymbol{\xi}_t\right)^2}{\sigma_{s,i}^2}\right) \quad (5)$$

where $x_{t,i}$ indicates the i-th element of the vector $\mathbf{x}_t$, $\mathbf{H}_s(i)$ denotes the i-th row vector of $\mathbf{H}_s$, $\boldsymbol{\xi}_t$ is a control vector at time t and $\sigma_{s,i}^2$ means the i-th element of the diagonal covariance matrix of the distribution s.

Setting the derivative of the objective function in (5) with respect to $\mathbf{H}_s(i)$ to zero, the solution is given by

$$\widehat{\mathbf{H}}_s(i) = \left(\sum_{t=1}^{T}\gamma_t(s)\frac{1}{\sigma_{s,i}^2}x_{t,i}\,\boldsymbol{\xi}_t^{'}\right) \times \left(\sum_{t=1}^{T}\gamma_t(s)\frac{1}{\sigma_{s,i}^2}\,\boldsymbol{\xi}_t\,\boldsymbol{\xi}_t^{'}\right)^{-1}. \quad (6)$$

Note that (6) is equivalent to the parameter estimate obtained in [5].

4. FACTORED MLLR

Let $\mathbf{M}$ and $\mathbf{b}$ depend on a control parameter η which is a d-dimensional continuous-valued vector. This implies that the mean vector of the distribution s is adapted differently depending on η. In this scenario, (1) is rewritten as

$$\widehat{\boldsymbol{\mu}}_s = \mathbf{M}(\eta)\,\boldsymbol{\mu}_s + \mathbf{b}(\eta) \quad (7)$$

and we call this approach the FMLLR technique.

In the FMLLR approach, $\mathbf{M}(\eta)$ and $\mathbf{b}(\eta)$ with diagonal assumption are represented as follows:

$$\mathbf{M}(\eta) = diag(\mathbf{w}_1^{'}\,\boldsymbol{\xi}, \mathbf{w}_2^{'}\,\boldsymbol{\xi}, \cdots, \mathbf{w}_p^{'}\,\boldsymbol{\xi}) \quad (8)$$

$$\mathbf{b}(\eta) = (\mathbf{v}_1^{'}\,\boldsymbol{\xi}, \mathbf{v}_2^{'}\,\boldsymbol{\xi}, \cdots, \mathbf{v}_p^{'}\,\boldsymbol{\xi})^{'} \quad (9)$$

where $\boldsymbol{\xi} = \phi(\eta)$ is a L-dimensional vector obtained by transforming the control parameter vector η using the function $\phi(\cdot)$. In (8) and (9), $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_p\}$ and $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_p\}$ are sets of L-dimensional regression vectors that are to be estimated.

Different from the conventional MLLR, the adaptation data have not only the speech features but also the corresponding control parameters. Let $(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T)$ be the adaptation vector sequence, $\boldsymbol{\xi}_t = \phi(\eta_t)$ be the transformed control parameter accompanied to $\mathbf{x}_t$. Then the likelihood function is defined and maximized as follows:

$$\{\widehat{\mathbf{W}}, \widehat{\mathbf{V}}\} = \arg\max_{\mathbf{W},\mathbf{V}} \mathcal{L}(\mathbf{W}, \mathbf{V}) \quad (10)$$

where

$$\mathcal{L}(\mathbf{W}, \mathbf{V}) = -\frac{1}{2}\sum_{t=1}^{T}\gamma_t(s)\left(\sum_{i=1}^{p}\frac{(x_{t,i} - \mathbf{w}_i^{'}\,\boldsymbol{\xi}_t\mu_{s,i} - \mathbf{v}_i^{'}\,\boldsymbol{\xi}_t)^2}{\sigma_{s,i}^2}\right)$$

in which $\boldsymbol{\mu}_{s,i}$ denotes i-th element of the mean vector $\boldsymbol{\mu}_s$, $\hat{\mathbf{W}}$ and $\hat{\mathbf{V}}$ are the updated parameters. The solution to (10) is obtained by setting the gradients of $\mathcal{L}\left(\mathbf{W},\ \mathbf{V}\right)$ with respect to $\mathbf{W}$ and $\mathbf{V}$ to zero as follows:

$$\frac{\partial\mathcal{L}}{\partial\widehat{\mathbf{w}}_i} = -2\sum_{t=1}^{T}\gamma_t(s)\frac{1}{\sigma_t^2}\left(x_t - \hat{\mathbf{w}}_i^{'}\,\boldsymbol{\xi}_t\mu_{s,i} - \hat{\mathbf{v}}_i^{'}\,\boldsymbol{\xi}_t\right)\mu_{s,i}\,\boldsymbol{\xi}_t$$

$$\Rightarrow \left(\sum_{t=1}^{T}\gamma_t(s)\frac{\mu_{s,i}^2}{\sigma_{s,i}^2}\,\boldsymbol{\xi}_t\,\boldsymbol{\xi}_t^{'}\right)\widehat{\mathbf{w}}_i + \left(\sum_{t=1}^{T}\gamma_t(s)\frac{\mu_{s,i}}{\sigma_{s,i}^2}\,\boldsymbol{\xi}_t\,\boldsymbol{\xi}_t^{'}\right)\widehat{\mathbf{v}}_i = \left(\sum_{t=1}^{T}\gamma_t(s)\frac{x_{t,i}\mu_{s,i}}{\sigma_{s,i}^2}\,\boldsymbol{\xi}_t\right) \quad (11)$$

$$\frac{\partial\mathcal{L}}{\partial\widehat{\mathbf{v}}_i} = -2\sum_{t=1}^{T}\gamma_t(s)\frac{1}{\sigma_t^2}\left(x_t - \hat{\mathbf{w}}_i^{'}\,\boldsymbol{\xi}_t\mu_{s,i} - \hat{\mathbf{v}}_i^{'}\,\boldsymbol{\xi}_t\right)\boldsymbol{\xi}_t$$

$$\Rightarrow \left(\sum_{t=1}^{T}\gamma_t(s)\frac{\mu_{s,i}}{\sigma_{s,i}^2}\,\boldsymbol{\xi}_t\,\boldsymbol{\xi}_t^{'}\right)\widehat{\mathbf{w}}_i + \left(\sum_{t=1}^{T}\gamma_t(s)\frac{1}{\sigma_{s,i}^2}\,\boldsymbol{\xi}_t\,\boldsymbol{\xi}_t^{'}\right)\widehat{\mathbf{v}}_i = \left(\sum_{t=1}^{T}\gamma_t(s)\frac{x_{t,i}}{\sigma_{s,i}^2}\,\boldsymbol{\xi}_t\right). \quad (12)$$

The final form would be acquired from the combining (11) and (12). For more details, the readers are referred to [7].

5. EXPERIMENTS

In order to evaluate the performance of the proposed technique, we conducted experiments on objective measurement and subjective listening test for expressive speech synthesis. All the speech data collected for expressive speech synthesis were spoken Korean language. For the construction of reading-style speech synthesizer, we used the reading-style speech data spoken by two speakers: YMK and HNC. YMK was a female and HNC was a male speaker, and each speaker provided 4,000 utterances of reading-style speech data amounting to 525 and 507 minutes, respectively. A baseline reading-style speech synthesizer was trained based on these speech database for each gender separately. We also collected the expressive speech data from the utterance of two other speakers: JEK (male) and SKJ (female). These speakers pronounced each sentence in four different types of emotional conditions: angry, sad, joyful and fearful. A total of 254 utterances amounting to 21 minutes on average were spoken for each type of emotional state by each speaker. Each utterance was sampled at 16 kHz and a 20 ms Hamming window was applied with 5 ms frame shift for speech feature extraction. As for the spectrum feature, a 25th-order mel-scaled cepstrum vector was extracted at each frame. By attaching the Δ- and $\Delta\Delta$-cepstra derived from the extracted mel-scaled cepstrum sequence, the spectrum feature could be represented by a 75-dimensional vector at each frame. We also extracted the pitch from each frame for the generation of voiced excitation signals. As the basic unit of speech synthesis, we applied quinphones followed by context-dependent reading-style text analysis described in [8]. Each quinphone was modeled by a 5 state left-to-right structured HMM where the observation distribution at each state was given by a single Gaussian PDF with diagonal covariance matrix. In this experiment, we aimed at adapting the HMM parameters of the baseline reading-style speech synthesizer to the given expressive speech data. Since, in our experimental conditions, the speakers of the adaptation data are different from those of the baseline system, parameter adaptation performs reduction of the mismatch caused by not only the different emotional states but also speaker variability. The baseline reading-style speech synthesizers for the speakers YMK and HNC were trained based on the decision tree technique presented in [9], which resulted in 7,215 and 5,797 leaf nodes, respectively. Among the 254 utterances for each type of emotion, we used 200 utterances for training the regression matrices of each method and the remaining 54 utterances for evaluating the performances for each gender. The control vectors of the expressive speeches were set to (1, 1, 0, 0, 0), (1, 0, 1, 0, 0), (1, 0, 0, 1, 0) and (1, 0, 0, 0, 1) for the angry, sad, joyful and fearful states, respectively.

Table 1. Average cepstral distance for male voice

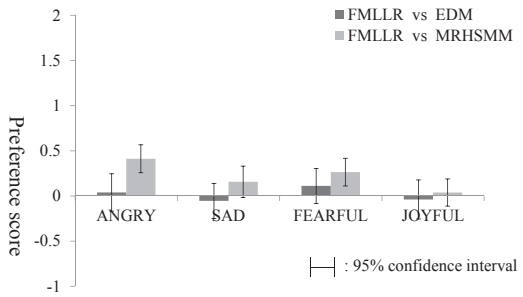
emotion \ method	EDM	MRHSMM	FMLLR
ANGRY	0.8651	1.0959	0.8830
JOYFUL	0.7407	0.9529	0.7417
FEARFUL	0.9548	0.9928	0.8514
SAD	0.7779	0.8954	0.7473

Table 2. Average cepstral distance for female voice

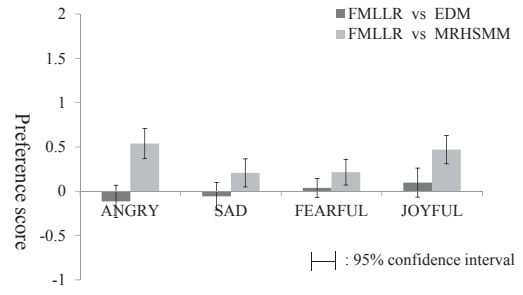
emotion \ method	EDM	MRHSMM	FMLLR
ANGRY	1.0606	1.1842	0.9834
JOYFUL	1.0988	1.1872	0.9980
FEARFUL	1.1256	1.1529	0.9659
SAD	0.9550	1.0480	0.8435

5.1. Objective performance evaluation for expressive speech synthesis

For the purpose of comparison, we also adapted the parameters of the reading-style speech synthesizer to the expressive speech using the adaptation method supported by HTS [9] and obtained the parameters of the expressive style speech synthesizer for each style and gender separately. The numbers of MLLR transform matrices for the expressive speech synthesizer of the speaker JEK were 531, 538, 511 and 682 for angry, fearful, joyful and sad, respectively. For the speaker SKJ, the numbers were 531, 415, 576 and 700, respectively. The resulted HMM parameters are referred to as the expression dependent models (EDMs). In the EDMs, not only the spectrum parameters but also the pitch and duration parameters were separately trained for each emotional state. For MRHSMM, the multiple regression matrix $\mathbf{H}_s$ was computed by (6). The numbers of $\mathbf{H}_s$ were 438 and 461 for SKJ and JEK, respectively. In the experiments, we adapted only the parameters associated to spectral features while the pitch and



(a) male



(b) female

Fig. 1. Result of subjective listening test.

duration were generated from the baseline synthesizer, i.e., EDM. For FMLLR, we trained the transforms along with (11) and (12) for each gender. For FMLLR, the numbers of transform matrices were maintained the same to those of the MRHSMM. Tables. 1 and 2 show the average cepstral distances between the actual and synthesized speech signals computed for the male and female speakers, respectively. From the results, it is seen that FMLLR outperformed other techniques. It is also interesting to find that the FMLLR algorithm produced better result even when compared with EDMs which is a matched training technique.

5.2. Subjective listening test for expressive speech synthesis

For this experiment, we synthesized 6 sentences for each expressive style by applying different methods, and 20 listeners participated for quality evaluation. EDMs, MRHSMM and FMLLR were evaluated in terms of paired comparison for each combination where for each test a pair of two speech files were given and the subject provided the relative quality of the latter file compared to the former in the range of [-3, 3] where positive value indicates that the former shows a better quality than the latter, and vice versa. To generate the synthesized speech file, the spectrum parameter was generated by each method while the pitch and duration parameters were set to the same to those of EDMs for each emotion. From the results shown in Fig. 1, we can confirm that the proposed approach produced a better quality of expressive speech than the conventional method.

6. CONCLUSIONS

In this paper, we have applied FMLLR to expressive speech synthesis as a novel technique for adapting the HMM parameters when the adaptation should depend on varying control parameters. The proposed approach provides compact parametric form to the MLLR framework, and the relevant parameters can be estimated according to the ML criterion. From the experimental results it has been found that FMLLR outperformed the MLLR and MRHSMM in terms of an objective measure as well as the subjective listening quality measure.

7. REFERENCES

[1] 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.

[2] B. Mak and R. Hsiao, "Kernel eigenspace-based MLLR adaptation", *IEEE Trans. Audio, Speech and Language Processing*, vol.15, no.3, pp.784-795, Mar. 2007.

[3] Z. Karam and W. Campbell, "A Multi-class MLLR kernel for SVM speaker recognition", *ICASSP*, Las Vegas, NV. pp.4117-4120, 2008.

[4] T. Nose, Y. Kato, and T. Kobayashi, "A speaker adaptation technique for MRHSMM-based style control of synthetic speech," *ICASSP*, Honolulu, HI, pp. 833-836, 2007.

[5] M. Tachibana, S. Izawa, T. Nose, and T. Kobayashi, Speaker and style adaptation using average voice model for style control in HMM-based speech synthesis, *ICASSP*, Las Vegas, NE, pp. 4633-4636, 2008.

[6] T. Nose, M. Tachibana, and T. Kobayashi, "HMM-based style control for expressive speech synthesis with arbitrary speaker's voice using model adaptation", *IEICE Trans. Inf. & Syst.*, vol.E92-D, 3, pp.489-497, Mar. 2009.

[7] N. S. Kim, J. S. Sung, and D. H. Hong, Factored MLLR adaptation, *IEEE Signal Processing Letters*, vol.18, no.2, pp. 99-102, Feb. 2011.

[8] J. S. Sung, D. H. Hong, K. H. Oh, and N. S. Kim, "Excitation modeling based on waveform interpolation for HMM-based speech synthesis," *Interspeech*, Makuhari, Japan, pp. 813-816, Sep. 2010.

[9] H. Zen et al., "The HMM-based speech synthesis system version 2.0," *Proc. of ISCA SSW6*, Bonn, Germany, Aug. 2007.