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

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

Bayesian Mixture of Probabilistic Linear Regressions for Voice Conversion

Na Li[1,2] and Yu Qiao[1,3]

[1]Shenzhen key lab of CVPR, Shenzhen Institutes of Advanced Technology, CAS, China
[2]Northwestern Polytechnical University, Xi'an, China
[3]The Chinese University of Hong Kong, Hong Kong, China
{na.li, yu.qiao}@siat.ac.cn

10.21437/Interspeech.2012-31

Abstract

The objective of voice conversion is to transform the voice of one speaker to make it sound like another. The GMM-based statistical mapping technique has been proved to be an efficient method for converting voices [1, 2]. In a recent work [3], we generalized this technique to Mixture of Probabilistic Linear Regressions (MPLR) by using general mixture model of source vectors. In this paper, we improve MPLR by considering a prior for the transformation parameters of linear regressions, which leads to Bayesian Mixture of Probabilistic Linear Regressions (BMPLR). BMPLR has the effectiveness and robustness of Bayesian inference. Especially when the number of training data is limited and the mixture number is larger, BMPLR can largely relieve the overfitting problem. This paper presents two formulations for BMPLR, depending on how to model noise in probabilistic regression function. In addition, we derive equations for MAP estimation of transformation parameters. We examine the proposed method on voice conversion of Japanese utterances. The experimental results exhibit that BMPLR achieves better performance than MPLR.

Index Terms: Bayesian linear regression, mixture of probabilistic regressions, voice conversion,

1. Introduction

Voice conversion (VC) is a task to modify the speaker characteristics of utterance without changing its linguistic meaning. VC offers a flexible way to synthesize various types of speech, and is receiving extensive research interests in the field of speech engineering. Since utterances of two speakers differ from each other in many aspects, such as speech rate, duration, pitch, formant frequencies and speaking style etc. An ideal VC technique should take account of all these aspects. However, this is difficult in practice, some of these features might be difficult to calculate and some are difficult to convert. Currently, many VC techniques focus on the mapping of spectral features, and only conduct simple modifications for prosody features, i.e. f0 [4, 1, 2, 5, 6].

The key problem in spectral mapping is to estimate a mapping function which transforms the spectral frame vector of source speaker to that of target speaker. Several techniques have been proposed for this task, such as code book mapping [4], artificial neural networks [6], Gaussian mixture model [1, 2, 5] etc. Among them, the GMM-based statistical conversion methods are probable the most popular ones. These methods use GMM to model the densities of spectral vectors of source speaker or the joint spectral vectors of source and target speakers. Then a linear regression is estimated for each component, and the final conversion function is a weighted summation of these linear regressions. One key advantage of GMM based conversion technique comes from its mixture nature, which allows it to deal with nonlinear transformation. In a recent work, we introduce Mixture of of Probabilistic Linear Regressions (MPLR) [3], which unifies previous GMM based conversion methods [1, 2]. MPLR not only provides intrinsic understanding of previous methods, but also indicates better methods for parameters estimation.

Bayesian linear regression (BLR) generalizes classical linear regression by taking account of a prior distribution of transformation parameters. In many problems, the conjugate priors with Gaussian forms are used due to formulation convenience and effectiveness. Compared with classical LR, BLR can release the overfitting problem when the number of training samples is limited, and can lead to methods to determine model complexity. Inspired by the success of BLR on various problems, we study a Bayesian treatment for Mixture of Probabilistic Linear Regressions (BMPLR). We introduce two formulations of BMPLR depending on how to model noise in regression function. We develop methods to estimate the optimal transformation parameters for both formulations in sense of maximum a posterior (MAP). We examine the proposed methods on voice conversion of Japanese utterances with ATR 503 corpus. The experimental results show that BMPLR always outperforms pervious MPLR.

2. Mixture of Probabilistic Linear Regressions

Many signal processing and speech problems can be reduced to find a mapping relation between two spaces. Given a set of training examples, the estimation of a mapping function can be seen as a regression problem in statistics [7, 8]. This section describes a short review on linear regression and Mixture of Probabilistic Linear Regressions [3].

Let x denote a source vector with dimensionality d, and y denote a target scalar [1]. The objective of regression is to estimate a regression/mapping function $y' = f(x, \theta)$, where θ represents regression parameters. Let $\{x_i, y_i\}_{i=1}^N$ denote a set of training data pairs, and set $X = [x_1, x_2, ..., x_N]$ and $Y = [y_1, y_2, ..., y_N]$. The relationship between regressor $f(x_i, \theta)$ and y_i is modeled by considering noise ϵ_i, $y_i = f(x_i, \theta) + \epsilon_i$. In this paper, we assume Gaussian noise, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

In Linear Regression (LR), regression function f has a lin-

[1]In many problems, the target can be a vector. For these problems, we can train a regression function for each dimension of the target vector.

ear form[7],

$$f(x,\theta) = Bx + b. \tag{1}$$

With argument vector $\hat{x}_i = [x_i^T, 1]^T$, Eq. 1 can be simplified to $y_i' = A\hat{x}_i$. Regression parameter $A = [B, b]$ is a row vector of dimension $d+1$. Eq. 1 indicates the following conditional distribution $p(y|x, A, \sigma^2) = \mathcal{N}(y|A\hat{x}, \sigma^2)$. Usually, parameters A can be estimated by minimizing the summation of differences $\sum_i \|y_i - f(x_i,\theta)\|^2$ or maximizing likelihood $\prod_i p(y_i|x_i)$. Both lead to the same solution for LR,

$$A^* = Y\hat{X}^T(\hat{X}\hat{X}^\top)^{-1}, \tag{2}$$

where $\hat{X} = [\hat{x}_1, \hat{x}_2, ..., \hat{x}_I]$ and '$^\top$' denotes matrix transpose. According to Gauss-Markov theorem [8], among all unbiased linear transformations, the optimal estimation of Eq. 2 has the minimum variance.

Although LR is simple and efficient, it cannot handle nonlinear mapping relations which exist in many problem such as voice conversion. To deal with this problem, Mixture of Probabilistic Linear Regressions (MPLR) [3] conducts a 'soft' division of space of x by using the posterior probabilities $p(k|x)$ of a mixture model $p(x) = \sum_k w_k p(x|k)$, where $w_k = p(k)$ is a prior of k-th mixture. Then, we estimate a local linear transformation $A_k\hat{x}$ for each mixture component. The final regression function is a weighted combination of all linear regressions,

$$f_{\text{MPLR}}(x) = \sum_{k=1}^{K} p(k|x) A_k \hat{x}. \tag{3}$$

Posterior $p(k|x)$ can be calculated with the Bayes' theorem,

$$p(k|x) = \frac{w_k p(x|k)}{\sum_j w_j p(x|j)}. \tag{4}$$

It is noted that conversion with Eq. 3 is different from the Mixtures of Linear Regression models (Chapter 14.5 [9]), where the weights for linear regressions are fixed for all training samples. MPLR can be seen as a general form of previous GMM based voice conversion techniques [7, 8]. And the formulation of MPLR leads to two parameter estimation methods, one is minimum squared error, and the other is minimum decomposed squared error [3].

3. Bayesian Analysis of MPLR

3.1. Bayesian Linear Regression

Compared with classical linear regression, Bayesian linear regression (BLR) introduces a prior on transformation parameters A (Chapter 3.3 [9]). Since $p(y|x, A, \sigma^2)$ has the exponential form of a quadratic form of A, the corresponding conjugate prior has a Gaussian form $p(A) = \mathcal{N}(A|m_0, S_0)$, with mean m_0 and covariance S_0. BLR estimates the posterior probability by,

$$\begin{aligned} p(A|Y, X, \sigma^2) &= \frac{p(Y|A, X, \sigma^2)p(A)}{p(Y, X, \sigma^2)} \\ &= \mathcal{N}(A|m_N, S_N), \end{aligned} \tag{5}$$

where mean $m_N^\top = S_N(S_0^{-1}m_0^\top + \sigma^{-2}\hat{X}Y^\top)$ and $S_N^{-1} = S_0^{-1} + \sigma^{-2}\hat{X}\hat{X}^\top$. If we consider an infinite prior $|S_0| \to \infty$, then mean m_N reduces to the classical maximum likelihood estimation of A. On the other hand, if the number of training samples $N = 0$, the posterior distribution reduces to prior $\mathcal{N}(A|m_0, S_0)$. For many problems, a zero-mean isotropic Gaussian is popularly used as prior, $p(A) = \mathcal{N}(A|0, \alpha^{-1}I)$. Then the parameters of corresponding posterior distribution are $m_N^\top = S_N\sigma^{-2}\hat{X}Y^\top$ and $S_N^{-1} = \alpha I + \sigma^{-2}\hat{X}\hat{X}^\top$. With these assumptions, the predictive distribution of new sample x is a Gaussian distribution,

$$\begin{aligned} p(y|Y, X, x, \alpha, \sigma^2) &= \int p(y|x, A, \sigma^2)p(A|Y, X, \alpha, \sigma^2)dA \\ &= \mathcal{N}(y|m_N\hat{x}, \sigma^2 + \hat{x}S_N\hat{x}^\top). \end{aligned} \tag{6}$$

3.2. Bayesian Mixture of Probabilistic Linear Regressions

It is well known that Bayesian inference of linear regression can avoid the over-fitting problem of maximum likelihood, and also can lead to automatic methods of determining model complexity [8, 9]. The effectiveness of BLR has been widely verified in various problems. Inspired by the success of BLR, we conduct a Bayesian treatment for Mixture of Probabilistic Linear Regressions (BMPLR).

Before using Bayesian analysis, we need to extend the conversion Eq. 3 into a probabilistic form $p(y|x, \theta)$ at first, where θ denotes model parameters. There are two methods to define $p(y|x, \theta)$, depending on how to model noise. Details are given as follows,

Single noise: The first method uses a single noise $\epsilon \sim \mathcal{N}(0, \beta^{-1})$ for the total conversion function,

$$y = \sum_{k=1}^{K} p(k|x) A_k \hat{x} + \epsilon. \tag{7}$$

In this way, $p(y|x, \theta)$ is a Gaussian distribution,

$$p(y|x, \theta) = \mathcal{N}(y|\sum_{k=1}^{K} p(k|x) A_k \hat{x}, \beta^{-1}). \tag{8}$$

Multiple noise: In the second method, we consider noise for each of the mixture components, $\epsilon_k \sim \mathcal{N}(0, \beta_k^{-1})$,

$$y = \sum_{k=1}^{K} p(k|x)(A_k\hat{x} + \epsilon_k). \tag{9}$$

Then $p(y|x, \theta)$ has a GMM form,

$$p(y|x, \theta) = \sum_k p(k|x)p(y|x, k, \theta), \tag{10}$$

$$p(y|x, k, \theta) = \mathcal{N}(y|A_k\hat{x}, \beta_k^{-1}). \tag{11}$$

For each A_k, we consider a Gaussian prior, $p(A_k) = \mathcal{N}(A|m_0^k, S_0^k)$. For notation convenience, we set $V_0 = \{m_0^k, S_0^k\}$, $A = \{A_k\}$, and $\beta = \{\beta_k\}$ in Eq. 9. In the next, we assume V_0 and β are given. In BMPLR, the posterior distribution of transformation parameters A is given by,

$$p(A|X, Y, \beta, V_0) = \frac{p(Y|X, A, \beta)p(A|V_0)}{p(X, Y, \beta, V_0)}. \tag{12}$$

3.2.1. MAP estimation of BMPLR

This subsection discusses maximum a posterior (MAP) estimation of A with Eq. 12. Since $p(X, Y, \beta, V_0)$ is independent of A, we have

$$\arg\max_A p(A|X, Y, \beta, V_0) = \arg\max_A p(Y|X, A, \beta)p(A|V_0). \tag{13}$$

The two terms in the left side can be further decomposed by,

$$p(A|V_0) = \prod_k p(A_k|V_0) = \prod_k \mathcal{N}(A_k|m_0^k, S_0^k), \quad (14)$$

$$p(Y|X, A, \beta) = \prod_i p(y_i|x_i, A, \beta). \quad (15)$$

Then we calculate the log likelihood of Eq. 13

$$\begin{aligned}\mathcal{L}(A) &= \ln p(Y|X, A, \beta)p(A|V_0) \\ &= \sum_k \ln \mathcal{N}(A_k|m_0^k, S_0^k) + \sum_i \ln p(y_i|x_i, A, \beta). \quad (16)\end{aligned}$$

There are two definitions of $p(y_i|x_i, A, \beta)$, by Eq. 8 and Eq. 10. We discuss the MAP estimation for them separately. For notation convenience, let $\gamma_i^k = p(k|x_i)$.

MAP with single noise (Eq. 8):

$$\begin{aligned}\mathcal{L}(A) = &-\frac{1}{2}\sum_k (A_k - m_0^k){S_0^k}^{-1}(A_k - m_0^k)^\top \\ &-\frac{\beta}{2}\sum_i (y_i - \sum_k \gamma_i^k A_k \hat{x}_i)^2 + \text{const.} \quad (17)\end{aligned}$$

The above function has a quadratic form of A_k which can be solved directly. Let $\frac{\partial \mathcal{L}(A)}{\partial A_k} = 0$, we have K equations,

$$\begin{aligned}&A_k {S_0^k}^{-1} + \beta \sum_j A_j \left(\sum_i \gamma_i^j \gamma_i^k \hat{x}_i \hat{x}_i^\top\right) \\ &= m_0^k {S_0^k}^{-1} + \beta \sum_i \gamma_i^k y_i \hat{x}_i^\top. \quad (18)\end{aligned}$$

We can obtain optimal A_k^* by solving these K linear equations together.

MAP with multiple noise (Eq. 10):

$$\begin{aligned}\mathcal{L}(A) = &-\frac{1}{2}\sum_k (A_k - m_0^k){S_0^k}^{-1}(A_k - m_0^k)^\top + \\ &\sum_i \ln\left(\sum_k \gamma_i^k \mathcal{N}(y_i|A_k\hat{x}_i, \beta_k^{-1})\right) + \text{const.} \quad (19)\end{aligned}$$

To directly maximize Eq. 19 is difficult. Instead, we introduce the following approximate method which optimizes its upper boundary. Remind $\gamma_i^k \geq 0$ and $\sum_k \gamma_i^k = 1$. According to Jensen's inequality, we have

$$\ln\left(\sum_k \gamma_i^k \mathcal{N}(y_i|A_k\hat{x}_i, \beta_k^{-1})\right) \leq \sum_k \gamma_i^k \ln \mathcal{N}(y_i|A_k\hat{x}_i, \beta_k^{-1}). \quad (20)$$

With this inequality, we can introduce the following upper boundary of $\mathcal{L}(A)$ as,

$$\begin{aligned}\mathcal{L}'(A) &= -\frac{1}{2}\sum_k (A_k - m_0^k){S_0^k}^{-1}(A_k - m_0^k)^\top \\ &\quad -\sum_i \sum_k \frac{\gamma_i^k \beta_k}{2}(y_i - A_k\hat{x}_i)^2 + \text{const} \\ &= -\frac{1}{2}\sum_k ((A_k - m_0^k){S_0^k}^{-1}(A_k - m_0^k)^\top \\ &\quad + \sum_i \gamma_i^k \beta_k (y_i - A_k\hat{x}_i)^2) + \text{const.} \quad (21)\end{aligned}$$

Sine posterior probabilities of GMM $\{\gamma_i^k\}$ are sparse (one is near to 1; the others are near to zero), the above boundary is usually tight. An advantage of Eq. 21 is that $\max_A \mathcal{L}'(A)$ can be further decomposed into K optimization problem,

$$\arg\max_{A_k}(A_k - m_0^k){S_0^k}^{-1}(A_k - m_0^k)^\top + \beta_k \sum_i \gamma_i^k (y_i - A_k\hat{x}_i)^2. \quad (22)$$

The above equation can be solved directly. The optimal solution is given by,

$$A_k^* = (m_0^k {S_0^k}^{-1} + \beta_k \sum_i \gamma_i^k y_i \hat{x}_i^\top)({S_0^k}^{-1} + \beta_k \sum_i \gamma_i^k \hat{x}_i \hat{x}_i^\top)^{-1}. \quad (23)$$

It is noted that solving Eq. 22 is much more efficient than solving the equation sets Eq. 18, which need to optimize all $\{A_k\}$ at the same time and need to solve a larger scale linear equations.

4. Experiments

We evaluate the performance of Bayesian MPLR (BMPLR) on voice conversion tasks, and make comparisons with the classical MPLR methods. In [3], we develop two methods for estimating transformation parameters of MPLR. One is direct optimization (MPLR), the other is decomposed approximate optimization (MPLR-D). For BMPLR, we have proposed two methods for estimating parameters in Section 3. One is MPLR with single noise (MPLR-S) by solving Eq. 18, and the other is MPLR with multiple noise (MPLR-M) by solving Eq. 22. For the second method, it is difficult to optimize it directly. We use the approximate optimal solutions given by Eq. 23, which works well in practice.

One basic problem for using BMPLR on voice conversion is how to set prior distributions $\{p(A_k)\}$. For simplicity, we assume all these distributions are equal, $p(A_1) = p(A_2) = ... = p(A_k)$ since we don't have special information on each of A_k. We initialize mean as zero and covariance matrix S_0^k as $S_0^k = \alpha^{-1}I$. These priors have been widely used in previous Bayesian analysis. Under these setups, the most vital factor which influences training, is ratio $\frac{\alpha}{\beta}$, which is set with cross validation in our experiments. It is noted that one might improve the results by considering better prior, but this is out of the discussion of this paper.

We used a ATR-503 phoneme balanced corpus pronounced by a male speaker and a female speaker for evaluation. The sampling frequency of each utterance is 16k Hz. We used 24 dimension cepstral vector for speech representation, and converted the female voice to the male voice. The cepstrum distortion [1] between the target cepstral vector $[y_t^1, ..., y_t^{24}]$ and the converted cepstral vector $[y_c^1, ..., y_c^{24}]$ is used as an evaluation measure, $\text{CD[dB]}(y_c, y_t) = 10/\ln 10\sqrt{2\sum_{d=1}^{24}(y_t^d - y_c^d)^2}$.

We conduct two types of experiments for evaluation. To be fair for MPLR and BMPLR, we fix source GMM model $p(x)$ (with full covariance matrix) for all our experiments. In the first experiment, we use 10 utterances for training and another 31 utterances for testing. The mixture numbers are changed from 1 to 20. When the mixture number is 1, MPLR reduces to classical linear regression and BMPLR reduces to Bayesian linear regression. The results are shown in Fig. 1. Both BMPLR-S and BMPLR-M outperform MPLR based methods in all the setups. Because the number of training utterances is limited in this experiment, MPLR archives its best performance when

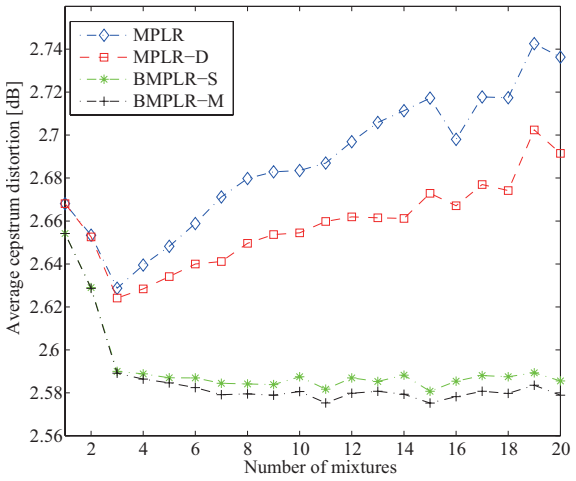

Figure 1: Voice conversion results of BMPLR and MPLR: average cepstrum distortion vs. mixture number.

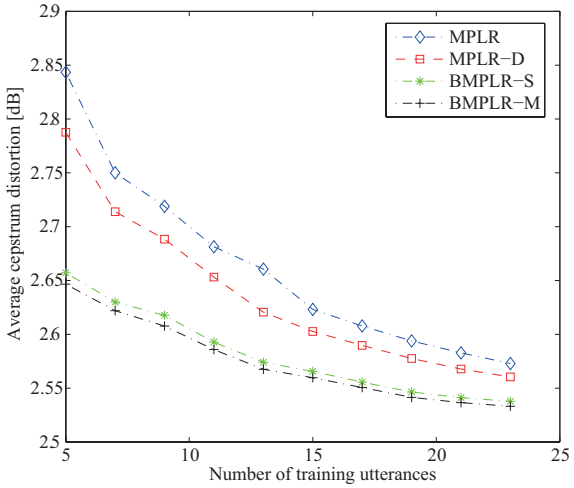

Figure 2: Voice conversion results of BMPLR and MPLR: average cepstrum distortion vs. training utterance number.

the mixture number is 3. Then the performance of MPLR and MPLR-M degrades as the mixture number increases. This indicates that MPLR and MPLR-M can overfit the training data when the mixture number is large and the model is complex. On the other hand, we find the average cepstrum distortion (ACD) of the proposed BMPLR-S and BMPLR-M exhibits a decreasing tendency as the mixture number increases. These results indicates that the proposed Bayesian methods can largely relieve the overfitting problems in MPLR, especially when the number of training utterance is limited and the mixture number is large.

In the second experiment, we fix the mixture number as 10, and change the number of training utterance as 4, 6,..., 24. The test utterances are set as the same as in the first experiment. As shown in Fig. 2, ACD of all the methods decreases as the number of training utterances increases. The proposed two methods BMPLR-S and BMPLR-M always achieve less ACD than previous MPLR methods. When the training data is sparse, our BMPLR based methods achieve significant improvements compared with MPLR based methods. In most cases, the BMPLR-M method outperforms the BMPLR-S method a bit but not significant. Remind solving Eq. 22 (BMPLR-M) is computationally efficient than solving Eq. 18 (BMPLR-S). We recommend to use BMPLR-M optimized with Eq. 22 in practice due to its effectiveness and efficiency.

5. Conclusions

In this paper, we propose a Bayesian treatment of Mixture of Probabilistic Linear Regressions (BMPLR) by introducing prior distribution of transformation parameters. We derive two types of formulations for BMPLR, by using different modeling of noise. The first formulation considers single noise for the total conversion, which leads to closed form solution for MAP parameter estimation. The second class considers multiple noise, which does not have closed form solution in MAP estimation. We develop a fast and approximate algorithm, which exhibits good results in experiments. Compared with previous MPLR methods, BMPLR has the advantages of Bayesian analysis and can relieve overfitting problems. Our experimental results show that BMPLR always outperforms MPLR in various experiments. Especially, when the number of training samples is small, BMPLR achieves significant improvements compared with MPLR.

6. Acknowledgement

This work is supported by National Natural Science Foundation of China (61002042), Shenzhen Basic Research Program for Distinguished Young Scholar (JC201005270350A), 100 Talents Programme of Chinese Academy of Sciences, and Introduced Innovative R&D Team of Guangdong Province "Robot and Intelligent Information Technology".

7. References

[1] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. on SAP*, vol. 6, no. 2, pp. 131–142, 1998.

[2] A. Kain and MW Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. ICASSP*, 1998.

[3] Y. Qiao and N. Minematsu, "Mixture of probabilistic linear regressions: a unified view of GMM-based mapping techniques," in *Proc. ICASSP*, 2009.

[4] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proc. ICASSP*, 1988.

[5] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on Gaussian mixture model withdynamic frequency warping of STRAIGHT spectrum," in *Proc. ICASSP*, 2001.

[6] S. Desai, A.W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Trans. on ASLP*, vol. 18, no. 5, pp. 954–964, 2010.

[7] D.C. Montgomery and E.A. Peck, *Introduction to linear regression analysis*, Wiley Series in Probability and Mathematical Statistics, 1982.

[8] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*, Springer New York, 2001.

[9] C.M. Bishop, *Pattern recognition and machine learning*, springer New York, 2006.