Regression-based Clustering for Hierarchical Pitch Conversion

Chung-Han Lee[1], Chi-Chun Hsia[2], Chung-Hsien Wu[1], and Mai-Chun Lin[1]

[1] Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan
[2] ICT-Enabled Healthcare Project, Industrial Technology Research Institute – South, Taiwan
Email: [1] { chlee, chwu, chun } @csie.ncku.edu.tw; [2] shiacj@itri.org.tw

ABSTRACT

This study presents a hierarchical pitch conversion method using regression-based clustering for conversion function modeling. The pitch contour of a speech utterance is first extracted and decomposed into sentence-, word- and sub-syllable-level features in a top-down mechanism. The pair-wise source and target pitch feature vectors at each level are then clustered to generate the pitch conversion function. Regression-based clustering, which clusters the feature vectors to achieve a minimum conversion error between the predicted and the real feature vectors is proposed for conversion function generation. A classification and regression tree (CART), incorporating linguistic, phonetic and source prosodic features, is adopted to select the most suitable function for pitch conversion. Several objective and subjective evaluations were conducted and the comparison results to the GMM-based methods for pitch conversion confirm the performance of the proposed regression-based clustering approach.

Index Terms: Regression-based, hierarchical, pitch conversion, clustering

1. INTRODUCTION

In order to eliminate the difficulty for collecting large speech databases, voice conversion, has been adopted as a post-process of expressive TTS to convert the synthesized neutral speech to expressive speech [1-3]. Recently, stochastic methods have dominated the development of voice conversion systems. The Gaussian mixture model (GMM) was applied to convert the source spectral features by a weighted sum of continuous conversion functions based on joint normality assumption [4]. Several improvements on GMM-based spectral conversion have been made by considering global variance [5] or dynamic features [2]. In prosody conversion, a codebook mapping method for pitch contour conversion [6] and the GMM- and HMM-based method has been adopted for spectral modification and joint spectral and pitch conversion [7][8].

In GMM-based voice conversion method, data samples, including source and target data, are clustered based on the distance to the centroid of each cluster. The prediction errors are defined as the vertical distances between the real target value and the predicted target value generated from the conversion function. Although clustering based on the distance to the centroid can minimize the quantization error, it cannot minimize the prediction distortion.

In this study, in order to reduce the prediction error, a regression-based clustering approach is proposed to cluster the data samples for the generation of conversion functions with minimum prediction distortion. To model pitch information, a hierarchical pitch structure is considered, in which the pitch contour of an input speech utterance is first extracted using the STRAIGHT algorithm [9] and the pitch contours were then represented by a set of Legendre polynomials. In pitch conversion, a CART is utilized to select the conversion function using linguistic and source prosodic information for pitch conversion. This study presents a hierarchical pitch conversion method for mandarin speech.

2. HIERARCHICAL PROSODY CONVERSION

From a top-down perspective, the pitch contour of a speech utterance can be decomposed into three pitch feature components at sentence-, word- and sub-syllable levels. For sentence level modeling, a simple, linear model is used.

$$g_0(t) = at + b \tag{1}$$

where $g_0(t)$ is the predicted pitch value at the sentence level at time t. Parameters a and b are estimated using the minimum mean square error (MMSE) criterion [10]. In the top-down procedure of the hierarchical pitch structure, the residuals of the pitch contour at the upper level are used as the curve fitting target of the lower level. At the word level, the discrete Legendre polynomials with order 3 are adopted to fit the pitch contour of a word as:

$$g_1(t) = P\left(\frac{n}{N}\right) = \sum_{i=0}^{2} a_i \Phi_i\left(\frac{n}{N}\right), \quad 0 \le n < N \tag{2}$$

where $g_1(t)$ is the predicted pitch value at the sentence level at time t with order 3. a_i and $\Phi_i(.)$ are the coefficients and the basis functions defined in [9] of the discrete Legendre polynomial, respectively. N is the length of pitch contour of a word. $P(.)$ is the original pitch contour.

At the sub-syllable level, the pitch contour of each tonal syllable can be divided into two parts and each part can be categorized and encoded according to the pitch values: high(H), middle(M) and low(L). For the five lexical tones in Mandarin which is a typical tonal language with different tone types in different morphemes, Tone 1(high pitch) can be encoded as HH, which represents a high pitch value in the first half part followed by another high pitch value in the second half part. Tone 2(rising pitch) is LH, Tone 3(low pitch) is LL, Tone 4(falling pitch) is HL, and the neutral tone is MM [11]. The coefficients of Legendre polynomials with order 4 are used to represent the pitch contour and estimated as:

$$a_i = \frac{1}{N+1} \sum_{n=0}^{N} P\left(\frac{n}{N}\right) \Phi_i\left(\frac{n}{N}\right), \quad 0 \le i \le 3 \tag{3}$$

In conversion function construction, the source and target feature vector sequences are denoted by a sequence of the aligned feature vector pairs $Z=\{z_1,z_2,\ldots,z_T\}$, where $\mathbf{z}_t=\left[\mathbf{x}_t',\mathbf{y}_t'\right]',t\in T$ (the total number of sub-syllables). For $X=\{x_1,x_2,\ldots,x_T\}$ and $Y=\{y_1,y_2,\ldots,y_T\}$, x_t and y_t are the source and target feature vectors (coefficients of Legendre polynomial), respectively, with the dimensionality equal to d. The distribution of Z is modeled as:

$$p(\mathbf{z}_t)=\sum_{m=1}^{M}w_m p(\mathbf{x}_t,\mathbf{y}_t\mid m)=\sum_{m=1}^{M}w_m N(\mathbf{z}_t;\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m) \quad (4)$$

where w_m are the prior probabilities of component m, and satisfies $\sum_{m=1}^{M}w_m=1$. M represents the total number of components. $N(\mathbf{z}_t;\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m)$ denotes the $2d$-dimensional Gaussian distribution with the mean vector $\boldsymbol{\mu}_m=\left[\left(\boldsymbol{\mu}_m^{X}\right)',\left(\boldsymbol{\mu}_m^{Y}\right)'\right]'$ and the covariance matrix $\boldsymbol{\Sigma}_m=\begin{bmatrix}\boldsymbol{\Sigma}_m^{XX} & \boldsymbol{\Sigma}_m^{XY}\\ \boldsymbol{\Sigma}_m^{YX} & \boldsymbol{\Sigma}_m^{YY}\end{bmatrix}$. The parameters $(w_m,\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m)$ for each component can be estimated using the EM algorithm [12]. The conversion function is then given by:

$$\begin{aligned}\tilde{\mathbf{y}}_t&=F(\mathbf{x}_t)=E[\mathbf{y}_t\mid\mathbf{x}_t]=\int_{\mathbf{y}_t}\mathbf{y}_t f(\mathbf{y}_t\mid\mathbf{x}_t)d\mathbf{y}_t=\int_{\mathbf{y}_t}\mathbf{y}_t\frac{f(\mathbf{y}_t,\mathbf{x}_t)}{f(\mathbf{x}_t)}d\mathbf{y}_t\\&=\int_{\mathbf{y}_t}\mathbf{y}_t\frac{\sum_i w_i f_i(\mathbf{y}_t,\mathbf{x}_t)}{\sum_i w_i f_i(\mathbf{x}_t)}d\mathbf{y}_t=\int_{\mathbf{y}_t}\mathbf{y}_t\sum_i\left[\frac{w_i f_i(\mathbf{x}_t)}{\sum_i w_i f_i(\mathbf{x}_t)}f_i(\mathbf{y}_t\mid\mathbf{x}_t)\right]d\mathbf{y}_t\\&=\sum_i\left[f(i\mid\mathbf{x}_t)\int_{\mathbf{y}_t}\mathbf{y}_t f_i(\mathbf{y}_t\mid\mathbf{x}_t)d\mathbf{y}_t\right]=\sum_i f(i\mid\mathbf{x}_t)E[\mathbf{y}_t\mid\mathbf{x}_t,i]\\&=\sum_{m=1}^{M}p(m\mid\mathbf{x}_t)\left[\boldsymbol{\mu}_m^{Y}+\boldsymbol{\Sigma}_m^{YX}\left(\boldsymbol{\Sigma}_m^{XX}\right)^{-1}\left(\mathbf{x}_t-\boldsymbol{\mu}_m^{X}\right)\right]\end{aligned} \quad (5)$$

where $p(m|x_t)$ represents the posterior probability of x_t belonging to component m, and is estimated as:

$$p(m\mid\mathbf{x}_t)=\frac{w_m N(\mathbf{x}_t;\boldsymbol{\mu}_m^{X},\boldsymbol{\Sigma}_m^{XX})}{\sum_{k=1}^{M}w_k N(\mathbf{x}_t;\boldsymbol{\mu}_k^{X},\boldsymbol{\Sigma}_k^{XX})} \quad (6)$$

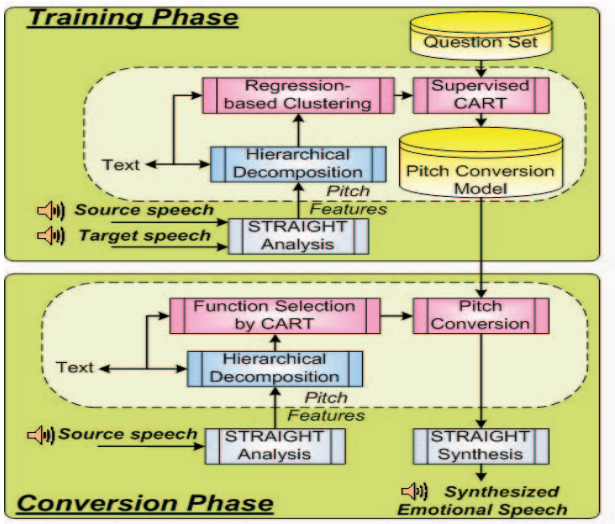

Figure 1: Flowchart of the proposed method.

Fig. 1 shows the flowchart of the proposed method. The pitch features of a sentence are decomposed into the sentence-, word- and sub-syllable-level features. The pitch features at different levels are converted separately using the conversion functions at the corresponding levels. Regression-based clustering and supervised CART is utilized to construct the pitch conversion model by incorporating the prosodic and linguistic features. In pitch conversion, the spectral and prosodic features of the input source speech are estimated by the STRAIGHT algorithm. The pitch features are hierarchically decomposed and converted using the conversion functions selected by CART. The STRAIGHT algorithm is used to synthesize the emotional speech by the converted spectrum and prosody.

3. REGRESSION-BASED CLUSTERING

The purpose of regression-based clustering is to cluster the data samples for the generation of conversion functions with minimum prediction distortion. Fig. 2 shows an example of two conversion errors for a data sample based on GMM-based and regression-based clustering methods, respectively. The four lines in these two figures represent four conversion functions obtained from the GMM-based and the regression-based clustering methods, respectively. The variation of the predicted target values using regression-based clustering is smaller than that for GMM-based clustering.

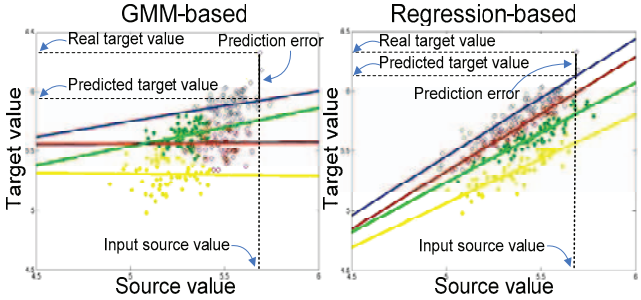

Figure 2: GMM-based vs. Regression-based clustering.

3.1 Clustering Algorithm

In regression-based clustering, a multi-dimensional linear regression model is adopted as the conversion function:

$$\mathbf{y}=f(\mathbf{x})=\boldsymbol{\beta}_0+\boldsymbol{\beta}_1\mathbf{x} \quad (7)$$

The parameter vectors (β_0, β_1) are estimated using MMSE criterion as:

$$\begin{aligned}E&=\sum_{i=0}^{n}[y_i-(\beta_0+\beta_1 x_i)]^2\\\frac{\partial E}{\partial\beta_0}&=-2\sum_{i=0}^{n}[y_i-(\beta_0+\beta_1 x_i)]=0\\\frac{\partial E}{\partial\beta_1}&=-2\sum_{i=0}^{n}[y_i-(\beta_0+\beta_1 x_i)]x_i=0\end{aligned} \quad (8)$$

The d-dimensional parameters of the linear regression

model are

$$\hat{\boldsymbol{\beta}}_1 = \frac{\sum\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\left(\mathbf{y}_i - \bar{\mathbf{y}}\right)}{\sum\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^2}, \quad \hat{\boldsymbol{\beta}}_0 = \bar{\mathbf{y}} - \hat{\boldsymbol{\beta}}_1 \bar{\mathbf{x}} \tag{9}$$

where $\bar{x}$ and $\bar{y}$ are the means of x and y, respectively. In regression-based clustering, multiple conversion functions are trained to further minimize the prediction errors (mean square errors). A similarity measure between a training sample S={**x**, **y**, **a**} and a cluster C is calculated as a combination of prosodic and linguistic similarities weighted by a power weight α (the value of α was determined as 0.2, 0.1 and 0.9 for anger, happiness and sadness emotions, respectively, from the experimental results):

$$Sim(C,S) = P_{Prosodic}(y \mid x, C)^{\alpha} \times P_{Linguistic}(a \mid C)^{1-\alpha} \tag{10}$$

where $P_{Prosodic}(y|x.C)$ denotes the prosodic similarity and is calculated as:

$$P_{Prosodic}(y \mid x, C) = N(y \mid \beta_1 x + \beta_0, \Sigma) \tag{11}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix. Prosodic similarity is estimated as an inverse proportion of the prediction error. The linguistic feature vector a=[a_1,a_2,…,a_l,…, a_L], and L is the total number of linguistic features. The linguistic similarity is calculated as:

$$P_{Linguistic}(a \mid C) = \left(\prod_{l=1}^{L} P_l(a_l \mid C)\right)^{1/L} \tag{12}$$

where $P_l(a_l|C)$ is the probability of linguistic feature a_l in cluster C. K-means algorithm is adopted for clustering. The parameters of K clusters are initialized (the value of K was determined to be 6, 4 and 4 for anger, happiness and sadness emotions, respectively, from the experimental results). The training samples are assigned to a cluster with the largest similarity, and the parameters of each cluster are recalculated. The process is repeated until no change occurs in cluster assignment.

3.2 Function Selection

Supervised CART is adopted to model the relation between the linguistic features and the clusters and is then used to retrieve an appropriate conversion function for pitch conversion. The following features are employed in the CART model:

- tone information (current, previous and following tones);
- phoneme information (previous final, current final/initial and following initial phoneme);
- word number (uni-gram, bi-gram, tri-gram, quad-gram)
- position in word (single, initial, medial, final);
- punctuation (： 、 ， 。 ？ ！);
- sub-syllable tone information (H, M, L).

In the training phase of CART, each sample is composed of cluster index and the corresponding features, including linguistic features and prosodic features of the source speech. Gain ratio is adopted as the splitting criterion and is estimated as:

$$GainRatio = \frac{\left(E_{parent} - \frac{N_{leftchild}}{N_{parent}} E_{leftchild} - \frac{N_{rightchild}}{N_{parent}} E_{rightchild}\right)}{SplitGains} \tag{13}$$

where E_{parent}, $E_{leftchild}$ and $E_{rightchild}$ denote the entropies of parent, left and right child nodes, respectively. The numerator of *GainRatio* represents the information gain for a split. The *SplitGains* represents the potential information generated by splitting a parent node with N_{parent} samples into two child nodes with $N_{leftchild}$ and $N_{rightchild}$ samples, respectively, and is calculated as:

$$SplitGains = -\frac{N_{leftchild}}{N_{parent}} \log \frac{N_{leftchild}}{N_{parent}} - \frac{N_{rightchild}}{N_{parent}} \log \frac{N_{rightchild}}{N_{parent}} \tag{14}$$

The split with the largest *GainRatio* in all possible splits is chosen and the tree growing stops when there is no significant information gain for all nodes.

4. EXPERIMENTS AND RESULTS

Experiments were designed and conducted to assess the performances for hierarchical pitch modeling, regression-based clustering and function selection using CART. For feature extraction, pitch contour and smoothed spectrum were extracted by the STRAIGHT algorithm. The analysis window was 23ms with a window shift of 8ms. Happiness, sadness and anger were adopted as the target emotions in this study. Three phonetically balanced small-sized parallel speech databases, each for one emotion, were designed and collected to train the voice conversion models. The numbers of sentences were 120, 110 and 115 for happiness, sadness and anger, respectively. The speaker was a female radio announcer, and was familiar with our study. All utterances were recorded at a sampling rate of 22.05 kHz and 16 bit resolution.

4.1 Comparison with GMM-based Clustering

For objective evaluation, mean square errors (MSE) was calculated between the predicted and the target pitch feature vectors as:

$$MSE = \frac{1}{M} \sum_{m=0}^{M-1} \left(\tilde{y}_m - y_m\right)^2 \tag{15}$$

where M is the total number of feature vectors. y_m and $\tilde{y}_m$ denote the target and the converted feature vectors.

To compare the GMM-based to the Regression-based pitch conversion methods, Fig. 3 shows the MSE of all the collected speech data in sub-syllable level as a function of mixture number for GMM-based and Regression-based clustering methods. The power weight α was set to 1.0 and only the conversion function of the mixture, which the input sample belongs to, was used. The MSE decreased with the increase of the cluster number and the decreasing rate of MSE was slower when the number of clusters exceeds 8. The MSE for sadness was potentially higher than happiness and anger. The analytical results indicate

that the regression-based clustering method has lower MSE than the GMM-based clustering for all emotions.

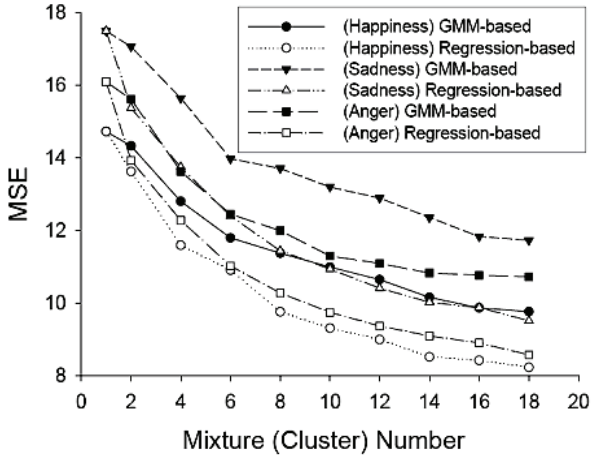

Figure 3: MSE as a function of mixture number

4.2 Evaluation of Formal Listening

The quality of the proposed method was also assessed during formal listening tests. Six listeners were asked to compare each converted sentence (with GMM-based or Regression-based method) with the sentences spoken by the target emotions of happiness, sadness and anger, respectively. A set of triads were presented to the listeners using the ABX method. X was either the converted speech by using the GMM-based method or the converted speech by using Regression-based method. A and B were either sentences spoken with the target (happiness, sadness and anger) or the source (neutral) emotion. The listeners were asked to select either A or B as being most similar to X. For the outside test, the ten fold cross validation is applied. Inside test means the test data is extracted from the training data. Table 1 presents the results from this test with the percentage of correct answers which means the converted speech was recognized as the target emotion. The proposed method with Regression-based method improves the accuracy of the conversion with the GMM-based method.

Table 1: ABX Listening test

	Inside		Outside	
Emotion	GMM-based	Regression-based	GMM-based	Regression-based
Anger	66%	91%	60%	82%
Happiness	60%	90%	56%	85%
Sadness	65%	70%	55%	59%

5. CONCLUSIONS

A hierarchical pitch conversion method using regression-based clustering has been presented in this study. The pitch contour of an input sentence is decomposed into sentence-, word- and sub-syllable-level features in a top-down procedure. The sample points are clustered to generate the conversion function with the minimum prediction error rather than the minimum distance to the centroid of a cluster. Prosodic similarity is estimated as an inverse proportion of prediction distortion. Linguistic similarity is integrated to improve the accuracy of function selection. CART is adopted to retrieve an appropriate conversion function using linguistic and source prosodic features. From the results, the proposed hierarchical pitch structure can reduce the variation in pitch modeling. The proposed regression-based clustering can effectively improve pitch conversion performance in MSE measure.

6. REFERENCES

[1] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari and K. Shikano, "GMM-based Voice Conversion Applied to Emotional Speech Synthesis," in *Proc. of EUROSPEECH'03*, pp. 2401-2404, Geneva, Switzerland, 2003.

[2] H. Duxans, A. Bonafonte, A. Kain and J. van Santen, "Including Dynamic and Phonetic Information in Voice Conversion Systems," in *Proc. of ICSLP 2004*, pp. 5-8, Jeju Island, South Korea, 2004.

[3] C. C. Hsia, C. H. Wu and J. Q. Wu, "Conversion Function Clustering and Selection Using Linguistic and Spectral Information for Emotional Voice Conversion," *IEEE Trans. Computers*, 56(9):1225-1254, 2007.

[4] Y. Stylianou, O. Cappé and E. Moulines, "Continuous Probabilistic Transform for Voice Conversion," *IEEE Trans. Speech and Audio Processing*, 6(2):131-142, 1998.

[5] T. Toda, A. W. Black and K. Tokuda, "Spectral Conversion based on Maximum Likelihood Estimation Considering Global Variance of Converted Parameter," in *Proc. of ICASSP 2005*, vol. 1, pp. 9-12, Philadelphia, USA, Mar 2005.

[6] O. Turk and L. M. Arslan, "Voice Conversion Methods for Vocal Tract and Pitch Contour Modification," in *Proc. of EUROSPEECH'03*, pp. 2845-2848, Geneva, Switzerland, 2003.

[7] T. En-Najjary, O. Rosec, and T. Chonavel, "A New Method for Pitch Prediction from Spectral Envelop and Its Application in Voice Conversion," in *Proc. of EUROSPEECH'03*, pp. 1753-1756, Geneva, Switzerland, 2003.

[8] C. H. Wu, C. C. Hsia, T. H. Liu, and J. F. Wang, "Voice Conversion Using Duration-Embedded Bi-HMMs for Expressive Speech Synthesis," *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 14, No. 4, July, 2006, pp.1109~1116.

[9] H. Kawahara, I. Masuda-Katsuse and A. de Cheveigné, "Restructuring Speech Representations using a Pitch Adaptive Time-Frequency-based F0 Extraction: Possible Role of a Repetitive Structure in Sounds," *Speech Communication*, 27(3-4):187-207, 1999.

[10] S. M. Kay, "Fundamentals of Statistical Signal Processing: Estimation Theory", *Prentice Hall PTR*, 1993.

[11] C. Huang, Y. Shi, J. Zhou, M. Chu, T. Wang and E. Chang "Segmental Tonal Modeling for Phone Set Design in Mandarin LVCSR," in *Proc. of ICASSP 2004*, vol. 1, pp. 901-904, Montreal, Canada, 2004.

[12] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. R. Statist. Soc. B*, vol. 39, pp. 1-38, 1977.