Review on GMM based Voice Transformation Techniques

Neha Yadav[1] Mr. Vinay Kumar Jain[2]
[1]M.E. Student [2]Associate Professor
[1,2]Department of Electronics & Telecommunication Engineering
[1,2]Shri Shankaracharya Technical Campus Bhilai, (C.G.)

Abstract— Voice Transformation is simply transformation or modification of voice from sound of one speaker into another specified speaker. This modifies the perceptual quality of the voice. During transformation quality as well as similarity of the voice is to be maintained which determines the efficiency of the transformation process. Transformation made is generally in the features like spectral mapping, frequency, amplitude etc. Key idea for transformation is the mapping of acoustical features. Many models such as GMM and RBF are developed to establish the mapping and obtain a transformation system. This paper describes many other methods on the basis of survey done. Mapping plays a key role in attaining transformation of the voice and quality is the challenge faced currently and it's the main point to be improved here.

Key words: GMM, RBF, MDSM, DFW, VQ, DWT

I. INTRODUCTION

Speech processing is one of the fields where large number of research is going on such as speech recognition, speech synthesis etc. One such area in speech processing is voice transformation. Voice transformation is the modification of voice produced by source speaker to be perceived by listeners as the voice of a different speaker the target speaker [1].VC IS Applied in area like films, action ,music industries, expressing emotion ,security, vocal pathology, voice restoration. By voice transformation quality of voice can be controlled. And to control voice transformation system are developed.

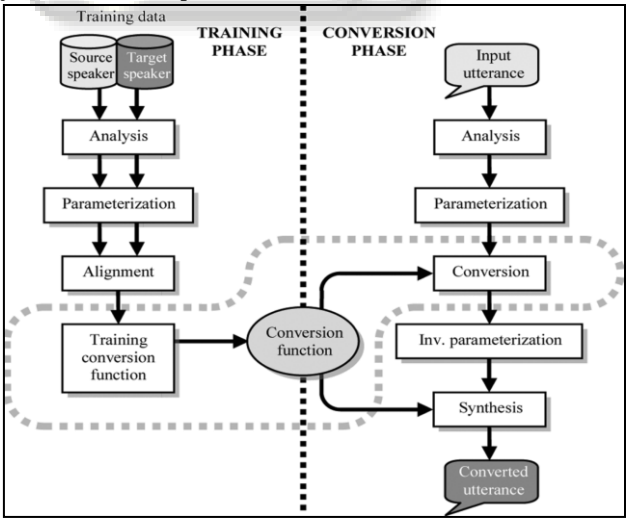

Fig. 1: Block diagram of a generic voice conversion system [1].

A voice conversion system contains training phase and conversion phase. During training phase, a conversion function is estimated from parallel source and target feature vector sequences. In conversion phase, the conversion function is applied on features extracted from new input speech of source speaker, and then the modified features are used to reconstruct the converted speech [2].

Basic applications such as speech to speech translation (SST) and text to speech translation (TTS). Basic idea is to map the acoustical features of sentences pronounced by a source speaker to values corresponding to the voice of target speaker[3].When the system are developed many problem occurs such as sometimes source and target speaker voice are not perfectly matched which makes the system inefficient. And this mismatch leads to over-smoothing.one of the other problem is with voice quality referred as muffled effect. Sometimes source voice itself is not noise free there are hissing noise, ringing noise, clicks etc which lowers the output quality.

To overcome such problem many algorithms have been developed in the past few years such as Vector quantization (VQ) and codebook sentences, Dynamic frequency warping (DFW),artificial neural network (ANN), Gaussian mixture model (GMM), Discrete Wavelet transform (DWT), Vocal tract length normalization (VTLN), Weighted frequency warping (WFW), Dynamic frequency warping plus amplitude scaling (DFWA), Bilinear frequency warping Plus amplitude scaling (BLFW), radial basis function (RBF), Correlation based Frequency warping (CFW), Minimum distance spectral mapping(MDSM). To improve quality and problems of standard GMM techniques many alternative methods are developed to improve quality of converted speech. Voice transformation using time scaling is successful but quality of pitch modification is still need improvement. Spectral mapping of source speaker to target speaker that is required for voice conversion are effective in transforming speaker identity but they produce low quality of voice [4]. Voice Transformation and Voice Conversion are based on the assumption that parallel databases are available for training purposes. However, there are many applications where parallel databases is impossible, like in cross-lingual voice conversion, where the source and target speakers speak different languages [5].On the basis of the surveying done, various standard GMM methods and their draw back and improved techniques studied are review.

II. LITERATURE SURVEY

A. Dynamic Frequency Warping of Straight Spectrum

The algorithm is proposed by Tomoki Toda et Al 2001.In this technique GMM is applied to the STRAIGHT (speech transformation and representation using adaptive interpolation of weighted spectrum)because GMM Based voice transformation techniques quality of converted speech degraded due to converted spectrum is exceedingly smoothed. By using dynamic frequency removes over smoothing introduced by GMM based voice transformation.

STRAIGHT is used in acoustical feature it is a high quality analysis synthesis technique based on pitch adaptive analysis combined with a surface reconstruction method in the time frequency region to remove signal periodicity .

To remove over-smoothing of converted speech use dynamic frequency warping performs spectral conversion and represented by warping function of original frequency axis and the converted frequency axis. After this warping function is represented as the path to minimized normalized spectrum distance between the STRAIGHT log spectrum of source and GMM based converted log spectrum [6].

To improve conversion accuracy on speaker we calculate GMM based converted log spectrum and the dynamic frequency warped log spectrum:

$$\|s_c(f)\| = \exp\left[\ln|s_d(f)| + w\langle \ln|s_g(f)| - \ln|s_d(f)|\rangle\right]$$
$$0 \le w \le 1 \qquad (1)$$

Where $s_d(f)$=dynamic frequency warped spectrum,
$s_g(f)$=GMM based converted spectrum respectively,
W= weight of residual spectrum.

To change all frequency necessary use not only weight of the constant value but also use frequency variant weights, change all frequency as follows:

$$w_h(f) = \begin{cases} \frac{2}{f_s} f & 0 \le f \le \frac{f_s}{2} \\ \frac{-2}{f_s} f + 2 & \frac{f_s}{2} \le f \le f_s \end{cases} \qquad (2)$$

$$w(f) = \begin{cases} -\frac{2}{f_s} f + 1 & 0 \le f \le \frac{f_s}{2} \\ \frac{2}{f_s} f - 1 & \frac{f_s}{2} \le f \le f_s \end{cases} \qquad (3)$$

Where f_s is sampling frequency if we use weight $w_h(f)$ increase frequency high then the converted spectrum is more closely to the GMM based converted spectrum.

B. Component Group GMM

The algorithm is proposed by Jianchun MA et Al 2005.CG-GMM frame work used for joint transformation of spectrum and pitch. Maximum likelihood estimation is used is used in CG-GMM [7].

CG-GMM is one a one element group and one component group, element of original D-dimensional feature vector X (as $X=[x_1, x_2 \dots x_p]$) split in *P* element group the dimensions of the element group is $x_1, x_2, \dots . x_p$ are $d_1, d_2 \dots d_p$ Satisfy $\sum_{i=1}^{p} d_i$, Component group $m_1, m_2 \dots m_p$ satisfy $\sum_{i=1}^{P} m_i = M$

Thus then component group Gaussian mixture model density function is given as;

$$p\left(\frac{X}{\lambda}\right) = \sum_{i=1}^{p} \sum_{j=1}^{m_i} \alpha_{ij} p_{ij}(x_i) \quad \dots (4)$$

$p_{ij}(x_i)$ Component density is dependent on component group the i^{th} group is equal to the corresponding element group is:

$$p_{ij}(x_i) = \frac{1}{(2\prod)^{\frac{d_1}{2}} |\Sigma_{ij}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\left(x_i - u_{ij}\right)' \sum_{ij} \left(x_i - u_{ij}\right)\right\}$$
$$i=1 \dots p \quad j=1 \dots m_i \qquad (5)$$

C. Viterbi Algorithm

The algorithm is proposed by Jian Zhi-Hua et Al 2007. Proposed Viterbi algorithm based on Gaussian mixture model in this method uses matrix of the transition probability of the target speaker to represent the relationship between sub discontinuities and high distortion in speech. The main purpose of Viterbi algorithm is conversion stage to find the optimal component of conversion function. Speech of given sentence is given by *(X1, X2, X3 …XT)* [8].

$$\{i_n^*, n = 1, \dots . T\} = \arg\max_{\{i_n, n=1, \dots T\}} [h(i_1, i2)\, h(i_2, x_2)$$
$$p(i_2, i_3) \dots . h(i_{T-1}, x_{T-1}) p(i_{T-1}, i_T) h(i_T, x_T) \qquad (6)$$

Where $\{i_n^*, n=1, \dots . T\}$= optimal component sequence for each source spectral vector.

$q(i_n, X_n)$=posterior probability and $p(i_{n-1}, i_n)$ is transmission probability.

The complete procedure to find viterbi algorithm of each source spectral vector is given by:

$$\widehat{y_n} = \mu_{i_n^*}^{y} + \sum_{i_n^*}^{yx} \left(\sum_{i_n^*}^{xx}\right)^{-1} . \left(X_n - \mu_{i_n^*}^{x}\right) \qquad (7)$$

D. Canonical Correlation Analysis Method

This algorithm proposed by ZhiHua Jian et Al 2007. Proposed canonical correlation method based on Gaussian mixture model. During conversion phase CCA consider variance of each component of the spectral vector Based on transforming the spectral characteristics of the source and target speaker. But not specify prosody modification .In Canonical correlation analysis method multivariate analysis focused on the amount of linear relationship between two sets of variables. It tries to find basis vectors for two sets of multidimensional variables such that the linear correlations between the projections onto these basis vectors are mutually maximized Suppose that the combination of two *l* -dimensional random vectors **x** and y The covariance matrix is given as:

$$\Sigma = \begin{bmatrix} \sum_{xx} & \sum_{xy} \\ \sum_{yx} & \sum_{yy} \end{bmatrix} \qquad (8)$$

Let consider two linear transformation $\eta = a^T x$ and $\Psi = b^T y$ vector a and b are correlation between η and Ψ is maximized as:

$$\rho = \max_{(a,b)} \frac{E(\eta\Psi)}{\sqrt{E(\eta^2)E(\Psi^2)}}$$
$$= \max_{(a,b)} \frac{a^T \sum_{xy} b}{\sqrt{(a^T \sum_{xx} a) \times (b^T \sum_{yy} b)}}$$
$$= \max_{(a,b)} (a^T \sum_{xy} b) \qquad (9)$$

The parameters a, b and ρ found by solving
The Eigen value equations:

$$\begin{cases} \sum_{xx}^{-1} \sum_{xy} \sum_{yy}^{-1} \sum_{yx} a = \rho^2 \ a \\ \sum_{yy}^{-1} \sum_{yx} \sum_{xx}^{-1} \sum_{xy} b = \rho^2 \ b \end{cases} \qquad (10)$$

Where a and b are the eigenvectors of the matrices $\Sigma^{-1}{}_{xx}\Sigma_{xy}\Sigma^{-1}{}_{yy}\Sigma_{yx}$ and $\Sigma^{-1}{}_{yy}\Sigma_{yx}\Sigma^{-1}{}_{xx}\Sigma_{xy}$ respectively, which correspond the same Eigen value We order the l Eigen values $\rho^2{}_1 \ge \rho^2{}_2 \ge \dots \dots \ge \rho_l{}^2$ and calculate the corresponding eigenvectors $\{a_1, a_2, \dots a_l\}$and $\{b_1, b_2, \dots . bl,\}$ Here a_1 and b_1 are called the first canonical variates, and ρ_1 is called the first canonical correlation[9].

E. Temporal Decomposition Analysis

The algorithm is proposed by Binh phu Nguyen et Al. 2008. TD analysis used for reducing drawback of over smoothing problem of GMM based voice conversion technique. In this method we need to modify the speech spectra of event target and event function not required to modify speech spectra frame-by-frame.

Temporal decomposition of speech into a sequence of overlapping target functions and corresponding event targets, given as [11]:

$$\hat{y}(n) = \sum_{k=1}^{k} a_k \emptyset_k(n), \quad 1 \le n \le N \quad \dots . (11).$$

Where a_k is the speech parameter corresponding to the k[th] event target.$\phi_k(n).\hat{y}(n)$ is n[th] spectral Parameter vector *y(n)* produced by TD model. *N, K* represents number of frames in the speech segment and number of event function. Limitation of the TD method is computational cost is high and parameter sensitivity is high for number and location of events. For reducing this drawback used modified restricted temporal decomposition technique (MRTD).

F. Artificial Neural Network

This algorithm proposed by Srinivas desai et Al 2009. ANN based voice transformation technique use parallel set of utterance for source and target speakers. For mapping of source speaker's spectral features to target speaker's spectral feature automatically extracts the relevant training data. ANN models consist of interconnected processing nodes, where each node represents the model of an artificial neuron, and the interconnection between two nodes has a weight associated with it. ANN models with different topologies perform different pattern recognition tasks. A multi-layer feed forward neural network is used in this work to obtain the mapping function between the input and the output vectors [12].

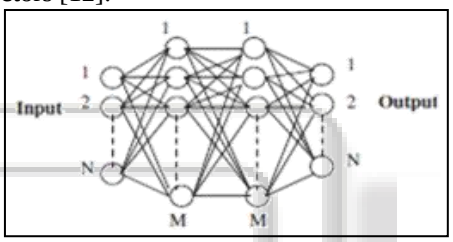

Fig. 2: Architecture of four layered ANN with N input and output nodes and M nodes in the hidden layer [12].

A radial basis function (RBF) neural network is used in this work to obtain the mapping function between the source and the residual vectors. It maintains nonlinear relationship between vocal tract and glottal excitation of speech frame [13].The radial basis function G(x) can be of various kinds. Typical choices are Gaussian, cubic, sigmoidal functions. Here, a Gaussian function is used [14]:

$$G(x) = \exp\left\{-\frac{\|x-c\|^2}{2\sigma^2}\right\} \qquad (12)$$

Where x,c,σ^2 are the input vector, the centre and variance of the radial basis function respectively output of RBF is given as:

$$y=\sum_{i=1}^{m} w_i G_i \qquad (13)$$

Where *Wi* are the output weights and m is the number of radial basis unit.

G. Weighted Frequency Warping

The algorithm is proposed by Daniel Erro et Al. 2010. WFW is extracted from GMM. It consists of time varying frequency warping function. it obtain good balance between similarity and quality. WFW method have following advantages: 1) it produces more natural converted voice compared to classical GMM method[1].2) it removes over smoothing problem.

To define frequency warping function $W_i(f)$ is an invertible function that maps the frequencies of the source speaker onto the target speaker. For an m^{th} order GMM, *m* different function W_i *(f)* is:

$$W^{(k)}(f) = \sum_{i=1}^{m} p_i\,(X^k).W_i(f) \qquad (14)$$

The probability that $x^{(k)}$ belongs to the i[th] Gaussian Component of the model $P_i(X^{(k)})$,$W_i(f)$ weighted coefficient The main advantage of the resulting frame-dependent frequency-warping function is that it does not contain significant temporal discontinuities.

H. Structured Gaussian Mixture Model

This is presented by Daojian Zeng et Al 2010. SGMM is applied to reduce the drawback of standard GMM techniques. SGMM techniques based on acoustical universal structure theory (ASU) applied in acoustic feature distribution in particular two speakers isolated.

According to the AUS theory, two speakers have the same distribution structure even if SGMM is trained separately with different corpus but containing sufficient balanced speech data in phonetics. The optimal aligned path between two structures can be carried out as follows:

Define distance between two SGMMs P and Q as Follows:

$$D(P, Q) = \sqrt{\frac{1}{n}\sum_{i<j}\left(p_{ij} - q_{ij}\right)^2} \qquad (15)$$

Here *i* and *j* are the nodes indexes, n is the number of nodes in structure *P* or *Q*. p_{ij} is the distance between nodes *i* and *j* in structure.

The source speaker structure P fixed and rotate the target speaker structure Q. When D(P,Q) reach to the minimum, the two structures are aligned in theory.

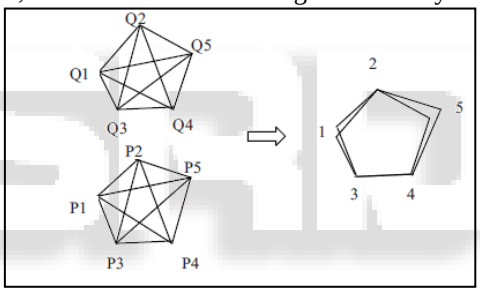

Fig. 3: Two SGMMs matching [16].

I. Dynamic Frequency Warping

This is presented by Elizabeth godoy et Al.2012.proposed dynamic frequency warping plus amplitude scaling that does not rely on GMM based transformation Specifically, instead of depending on a frame-by-frame alignment of the source and target speech, which proves inefficient in capturing a link that can be effectively exploited in transformation, DFWA functions on an acoustic-class level[16]. Advantage of DFW with amplitude scaling is that converted speech is more natural sound as compare to classical GMM method. It is good for cross language voice conversion. DFWA offers a more flexible framework for spectral envelope transformation that eliminates the need for parallel corpora in VC.

Frequency warping is based on smoothed histogram analysis of spectral peak occurrences in frequency. Histogram of the peak occurrence is carried out for each class q and frequencies of the histogram is separated by 50HZ for each class q estimation DFW function determine maxima of the source and target soothed histogram. for associated histogram maxima the DFW function for each class q is given as: $f \in [f^x_{q,im}, f^x_{q,im+1}]$.

$$W_q(f)=B_{q,m}f+C_{q,m} \qquad (16)$$

The warping function for a source frame x_n is represented as:

$$W(x_n, f)=\sum_{q=1}^{Q} w_q^x(x_n)\, W_q(f) \quad \ldots.(17).$$

This is the form of weighted frequency warping function $w_q^x(x_n)$. Corresponds to the mixture weight classifying the source frame according to the defined source feature space the source frame according to the defined source feature space.

Amplitude scaling is used after DFW estimation. AS is used to minimize error between the converted and target signals. Amplitude scaling function is estimated after DFW estimation using statics observed from the target and warped source data.

Let us consider $s^{y(n)}(f)$ is the target spectral envelope for n frame .then for class q the mean target log amplitude spectrum. The frames in learning set are given as:

$$\overline{\log(S_q^y(f))} = \sum_{n=1}^{N_y} w_q^y(y_n)\log(S^{y_n}(f)) \qquad (18)$$

Within a class q, the amplitude scaling function is intended to adjust the mean warped source log amplitude spectrum towards the mean target log amplitude spectrum. Thus, the amplitude scaling function for class *q*, $A_q(f)$ is given below[16]

$$\overline{\log(S_q^v(f))} = \sum_{n=1}^{N_x} w_q^x(x_n)\log(S^{v_n}(f)) \quad (19)$$

J. Bilinear Frequency Warping

This algorithm proposed by Daniel Erro et Al. 2013. Proposed bilinear frequency warping plus amplitude scaling. It is fully parametric (Cepstral) domain. It is referred as a bilinear FW+ AS (BLFW+AS) bilinear function are defined by one single parameter .Typically applied to perform vocal tract length normalization (VTLN). It performs conversion function at frame level thus it requires real time system. Bilinear frequency warping plus amplitude scaling succeeds at tackling over smoothing and at preserving the quality of the signals better than statistical methods [3].

Bilinear warping function requires only one parameter α it is defined as Z domain:

$$z_\alpha^{-1} = \frac{Z^{-1}-\alpha}{1-\alpha Z^{-1}}, \qquad |\alpha|<1 \quad (20)$$

By using p-dimensional cepstral vector, it has been proven that the cepstral vector that corresponds to the frequency warped version of the spectrum $p\times p$ *matrix* represented by is given by [16]:

$$y=W_{\alpha x,}W_{\alpha,} = \begin{bmatrix} 1-\alpha^2 & 2\alpha-2\alpha^2 & \ldots. \\ -\alpha+\alpha^3 & 1-4\alpha^2+3\alpha^4 & \ldots. \\ \vdots & \vdots & \ddots \end{bmatrix} \quad (21)$$

Mapping between original and warped frequencies is:

$$w_\alpha = \tan^{-1}\frac{(1-\alpha^2)\sin w}{(1+\alpha^2)\cos w-2\alpha} \qquad (22)$$

K. Corelation based Frequency Warping

This algorithm Proposed Xiaohai tain et Al.2014.CFW Based warping process used for calculating FW function on 513 dimensional spectral envelope. This method tried to remove optimization problem by modifying segment length .spectra are split in to segment and maximize segment correlation at the time of warping target segment are fixed and segment of source spectra will varied. It is improved method for speaker identity of converted speech and for quality of voice.

CFW warping process target segments boundaries are fixed while a source segment varies by maximizing correlation optimal warping path achieved. Given spectral envelopes are divided equal number of segments for source and target is written as:

$$\{(p_{s,1}^b\ p_{s,1}^e)\ldots\ldots(p_{s,N}^b, p_{S,N}^E)\}, \{(p_{t,1}^b, p_{t,1}^e)\ldots..(p_{t,N}^b, p_{t,N}^e)\} \quad (23)$$

The length of n[th] source and target segments given as $l_{s,n}$ And $l_{t,n}$ for correlation based frequency warping with source and target paired spectral envelope obtained by using backward step and tracking step.

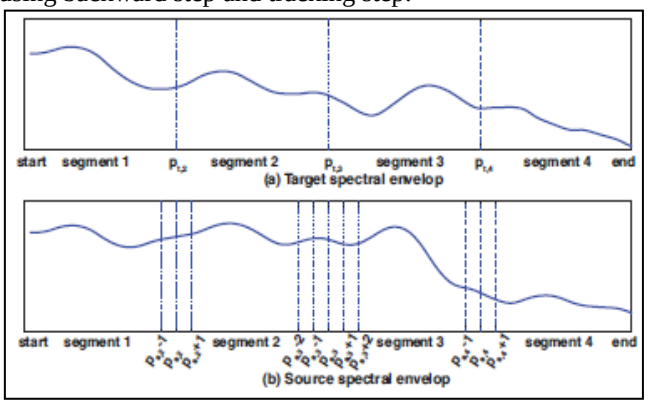

Fig. 4: An example of the procedure of correlation-base frequency warping. Figure (a) the target spectral envelope and predetermined boundaries. Figure (b) the source spectral envelope and the range of segment boundaries [2].

Backward step: fig shows the source and target spectral envelopes. Process start from last 4[th] segment from spectral to perform linear interpolation he source and target segment s_n,t_n denote n[th] segment .correlation coefficient s_n, t_n is:

$$\gamma(s_n,t_n)=\mathrm{cov}(s_n,t_n)\sqrt{var(s_n).var(t_n)} \quad (24)$$

Tracking step: In this step maximum cumulative correlation is selected. After selection, correlation of all segments is calculated.

L. Modified Gaussian Mixture Model

For removing drawback of standard GMM technique proposed modified Gaussian mixture model such as modulation spectrum based post filter (MSPF)[19].matrix variate GMM[20].target frame selection(TFS)[21].

GMM based voice conversion degrade the modulation spectrum of speech parameter. So we filter the generated parameter closely of the MS to natural speech in conversion stage. In MSPF based VC technique we generate naturally fluctuate temporal parameter sequence.

The modulation scale (MS) is defined by log-scale power spectrum of temporal sequence.

$$S(y)=[s(1)^T, \ldots\ldots\ , s(d)^T, \ldots.\ , s(D)^T]^T,$$
$$S(d)=[s_d(0)^T, \ldots\ s_d(d)^T, \ldots\ , s_d(D_s)^T]^T, (25)$$

Where sd (f) = f[th] MS of the d[th] dimension of the Parameter sequences.

$[y_1(d, \ldots., y_T(d)]^T$ = modulation frequency index

Ds = one half number of the DFT length.

For MS based post filtering, post filtering is applied training data with natural and converted data of target speaker's voice. During training phase probability distribution function is determine from natural speech [19].

$$P(s(y)/\lambda s)=N(s(y); \mu^{(N)}, \Sigma^{(N)}) \quad (26)$$

Where N (.;$\mu^{(N)}$,$\Sigma^{(N)}$)= Gaussian distribution mean of a mean vector.

Matrix variate GMM based joint vector of source and target is first constructed. Then the joint probability distribution function (PDF) is modeled. Whose rows and column vectors the function.

Joint density is modeled by an MV-GMM a new sequence is generated Z= [Z_1, Z_2…Z_n] for joint matrix Z_t is:

$$P(Z_t/\lambda^{(Z)}) = \sum_{m=1}^{M} w_m \, N_{mv}\,(Z_t\,; M_m\,, U_m\,, V_m\,) \quad (27)$$

Where Z_t= [x_t,y_t]∈R^{D*S}.

M. Minimum Distance Spectral Mapping plus Amplitude Scaling

This algorithm Proposed Gui Jin et Al.2015.Using DTW time alignment algorithm we propose Minimum distance spectral mapping algorithm based on point- to- point mapping that can be detecting without the formant positions source and target spectral envelope X (I) and Y (I) are stored in a set of n values, which can be represented as {x (1), ... , x (n)) and {y(i), ... ,y (n)} . We define D (i, j) as

$$D(i,j)=D(x(i),y(j))=\min\begin{cases} D(i-1,j)+\lambda d(i,j) \\ D(i,j-1)+\lambda d(i,j) \\ D(i-1,j-1)+\lambda d(i,j) \end{cases} \quad (28)$$

Where 1 ≤ i, j ≤ n, D (0, 0) = 0, λ is a penalty factor, generally Belongs to [√2, 2], and d (i,j) is defined as

$$d(i,j)=\sqrt{|log^2(x(i))-log^2(x(j))|} \quad (29)$$

To find optimal path from (1,1) to (n,n) minimize D(n,n) Dynamic programming is a best approach after point-to point mapping between mean spectral envelopes is estimated.

Using point-to-point mapping for warping process, it adjust the spectral tilt and formants position without destroying the formants structure and provides better result between speech quantity and identity similarity. After study various papers their advantages, limitations are summarized with the help of table:

Year	Author	Technique	Advantage	Limitations
2001	Tomoki[6]	DFW & Straight	remove signal periodicity	analysis &synthesis is very high
2005	Jianchun MA[7]	CG-GMM	Satisfactory speech quality and speaker identifiably	Cannot perform only one parameter.
2007	Jian[8]	Viterbi algorithm	Avoids Spectral discontinuities.	Can be used only in the conversion stage.
2007	Zhi Hua[9]	CCA	Better spectral conversion result.	Not allowed prosody modification.
2008	Bin phu [10]	TD&GMM	Smooth ness of converted speech is good	high computational cost
2009	Srinivas desai [12]	ANN	Improve quality & and naturalness of converted speech.	Only perform nonlinear mapping.
2010	Daniel [1]	WFW	Good balance between similarity and quality.	Not completely convert the source voice into the target voice.
2010	Daojian zeng [16]	SGMM	Equivalent speech quality &speaker individuality	create problem for finding optimal path.
2012	Elizabeth godoy [17]	DFWA	Achieves higher converted speech quality.	Spectral tilt accounted slowly in the amplitude scaling.
2013	Daniel erro[3]	BFWA	Preserving quality of signal.	BFWA is not superior compare to FWA.
2014	Xiao haitian[2]	CFW	Optimize the warping path.	Affected by segment boundaries.
2014	Dai suke Saito [20]	MVGMM	Effectively model characteristics and feature.	Use non target data introduce distortion.
2014	Shinnosuke Takamichi[19]	MSPF	Post filtering is good for quality.	higher modulation frequency is not better

Table 1: different techniques of voice transformation

III. EVALUATION

For evaluation of speech data two types of evaluation process used for voice conversion.1) Objective evaluation 2) subjective objective evaluation determine frame-to-frame accuracy in transformation and subjective evaluation determine quality of the converted speech and of its similarity to the target speech were conducted in formal listening tests[5].

A. Objective Evaluation

In objective evaluation we calculate MSE (mean squared error) to describe average frame-to-frame accuracy. Following equation is used to calculate MSE:

$$MSE=10\log\frac{\sum_{n=1}^{N}\sum_{i=1}^{24}(\widehat{y_n}\,(i)-y_n(i))^2}{\sum_{n=1}^{N}\sum_{i=1}^{24}(y_n(i))^2}$$

B. Subjective Evaluation

In subjective evaluation we conduct MOS (mean opinion score) test for speech quality and identity similarity [9].based on various papers the experimental result of quality MOS represented by following table:

Year	**Author**	**Techniques**	**Quality MOS**
2007	Errow[13]	WFW	3.27 (M-M) 3.00 (M-F) 3.60 (F-M) 4.20 (F-F)
2008	Shuang[21]	FW	3.48
2008	Binh Phu Nguyan[10]	TD& GMM	3.89(M-F) 3.67(F-M)
2008	Desai[22]	ANN	≈2.70
2009	Srinivas Desai[12]	ANN	3.06(M-M) 3.0(M-F)

2010	Daojian Zeng[16]	SGMM	(Ex.1,Ex2) 3.9,3.5(M-M) 3.7,3.6(M-F) 3.4,3.1(F-M) 3.6,3.3(F-F)
2014	Xiaohai Tian[2]	(DFW+AS CFW+AS) (AMF+AS, CFW+MS)	21.5(±5.85) 78.5(±5.85) 49(±6.82) 51(±6.82)
2015	Kevin D'souza[26]	GMM	3.1667(M-M) 3(M-F)&4(F-M) 3.333(F-F)

Table 2: Experimental result for quality MOS in voice conversion system.

In the test of speaker individuality, an ABX test was conducted. A represents the source speaker, B represents the target speaker, and X represents the converted speech, which supplied from each one of the three test systems [29].based on survey using different techniques experimental result of ABX indices are shown below table:

Year	Author	Techniques	ABX index
2001	Toda[6]	GMM DFW	77% (M -M) 83% (F- F)
2005	Jianchun MA[7]	CG-GMM	$\approx 60-65\%$
2007	Jian Zhi-Hua[8]	Viterbi based	81.8%(M-M) 93.3%(M-F) 84.2%(F-F) 92.5%(F-M)
2007	ZhiHua Jian[9]	CCA	(within gender) 85.5%(M-M) 86.3%(F-F) (across gender) 94.1%(M-F) 93.4%(F-M)
2008	Binh Phu Nguyen[10]	TD & GMM	4.50(M-F)& 4.00(F-M)
2010	XieChen[25]	GMM	100%(M-F) 100%(F-M) 80.7%(F-F)

Table 3: Experimental results for ABX indices in voice conversion system

IV. Conclusion

In this paper various voice transformation method are reviewed and studied to improve quality of speech. We reviewed GMM based voice transformation techniques and identified problems of classical GMM techniques, problems occur due to modify any features of voice signals. To improve quality of classical GMM techniques use new warping function, point-to-point spectral mapping between source and target speaker plus amplitude scaling, provide better result compare to standard GMM .new warping function such as DFW, CFW, WFW, BFW provides high quality converted speech and preserve speech signal.

Reference

[1] Daniel Erro, Asunción Moreno, and Antonio Bonafonte "Voice Conversion Based on Weighted Frequency Warping" IEEE Transactions on audio, speech, and language processing, vol. 18, no. 5,pp922-931 July 2010.
[2] Xiaohai Tian, Zhizheng Wu, S. W. Lee, and Eng Siong Chng "Correlation-based Frequency Warping for Voice Conversion" 9th international symposium on chiness spoken language (ISCSLP) pp 211-215 IEEE2014.
[3] Daniel Erro, Eva Navas, and Inma Hernaez "Parametric Voice Conversion Based on Bilinear Frequency Warping Plus Amplitude Scaling" IEEE transcations on audio, speech, and language processing, vol. 21,no. 3, pp556-566march 2013.
[4] Yannis stylianou "voice transformation: A survey" pp3585-35882009.
[5] "Introduction to the special section on voice transformation" IEEE transaction on audio, speech, and language processing, vol. 18, NO. 5, pp 909-911 JULY 2010.
[6] Tomoki Toda,Hiroshi Saruwatari,Kiyohiro Shikano " Voice Conversion Algorithm on Gaussian Mixture Model With Dynamic Frequency Warping of STRAIGHT Spectrum".ICASSP pp127-130, 2001.
[7] Jianchun MA, Wenju LIU "Voice Conversion based on Joint Pitch and Spectral Transformation with Component Group-GMM" 2005.
[8] Jian Zhi-Hua, YANG Zhen "Voice conversion using Viterbi algorithm based on Gaussian mixture Model" Proceedings of 2007 International Symposium on Intelligent Signal Processing and Communication Systems pp-32-35Nov.28-Dec.1, 2007
[9] ZhiHua Jian Zhen Yang "Voice Conversion Using Canonical Correlation Analysis Based on Gaussian Mixture Model Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing. PP 210-215. 2007.
[10] Binh Phu Nguyen and Masato Akagi "Phoneme-based Spectral Voice Conversion Using Temporal Decomposition and Gaussian Mixture Model" IEEE pp 224-229 2008.
[11] B. S. Atal, "Efficient coding of LPC parameters by temporal decomposition, Proc. ICASSP, pp. 81–84, 1983.
[12] S. Desai, E. Veera Raghavendra, B. Yegnanarayana ,Alan W Black, Kishore Prahallad "Voice conversion using artificial neural networks" pp 3893-3896 ICASSP 2009.
[13] Jagannath Nirmal, Pramod Kachare, Suprava Patnaik, Mukesh Zaveri "Cepstrum Liftering based Voice Conversion using RBF and GMM" International conference on Communication and Signal Processing, pp570-575 April 3-5, 2013, India.
[14] Danwen Peng, Xiongwei Zhang, Jian Sun "Voice Conversion Based on GMM and Artificial Neural Network"pp1121-1124 IEEE2010.
[15] Daniel Erro, Asuncion Moreno "Weighted Frequency Warping for Voice Conversion" Interspeech, 2007.
[16] Daojian Zeng, Yibiao Yu "Voice Conversion Using Structured Gaussian Mixture Model" ICSP, pp 541-5442010 proceeding.
[17] Elizabeth Godoy, Olivier Rosec, and Thierry Chonavel "Voice Conversion Using Dynamic Frequency Warping with Amplitude Scaling, for Parallel or Nonparallel

Corpora" IEEE transactions on audio, Speech, and language processing, vol. 20, no. 4, pp 1313-1323may 2012.
[18] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," IEEE Trans. Speech Audio Process., vol. 13, no. 5, pp. 930–944, Sep. 2005.
[19] Shinnosuke Takamichi, Tomoki Toda, Alan W Black and Satoshi Nakamura "Modulation Spectrum-Based Post-Filter for GMM-Based Voice Conversion" APSIPA pp2014.
[20] Daisuke Saito, Hidenobu Doi, Nobuaki Minematsu, Keikichi Hirose "Voice conversion based on matrix variate Gaussian Mixture model pp 567-571ICSP2014 Proceedings.
[21] Hung-Yan Gu, Sung-Fung Tsai "Improving Segmental GMM Based Voice Conversion Method with Target Frame Selection"9th International Symposium on Chinese Spoken Language Processing (ISCSLP) pp 483-487, 2014.
[22] Gui Jin, Michael T. Johnson, Jia Liu, Xiaokang Lin "Voice Conversion Based on Gaussian Mixture Modules with Minimum Distance Spectral Mapping" IClST2015.
[23] Z. Shuang, R. Bakis, and Y. Qin, "IBM voice conversion Systems for 2007 TC-STAR evaluation," Tsinghua Science & Technology, vol. 13, no. 4, pp. 510–514, 2008.
[24] S. Desai, E. Raghavendra, B. Yegnanarayana, A. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in IEEE WSLT, 2008.
[25] Xie Chen, Wei-Qiang Zhang, Jia Liu, Xiuguo Bao "An Improved Method for Voice Conversion Based on Gaussian Mixture Model" 2010 International Conference on Computer Application and System Modeling (ICCASM),vol.4,pp404-407,2010.
[26] Kevin D'souzaK.T.V Talele "Voice Conversion Using Gaussian Mixture Models" International Conference on Communication, Information & Computing Technology (ICCICT), Jan. 16-17, 2015.
[27] Anderson F. Machado, Marcelo Queiroz "Voice conversion: A critical survey" 2010.

