Sequential Voice Conversion Using Grid-Based Approximation

Hadas Benisty, David Malah, and Koby Crammer
Department of Electrical Engineering
Technion, Israel Institute of Technology
Technion City, Haifa 32000, Israel
{hadasbe@tx,malah@ee,koby@ee}.technion.ac.il

***Abstract*—Common voice conversion methods are based on Gaussian Mixture Modeling (GMM), which requires exhaustive training (typically lasting hours), often leading to ill-conditioning if the dataset used is too small. We propose a new conversion method that is trained in seconds, using either small or large scale datasets. The proposed Grid-Based (GB) method is based on sequential Bayesian tracking, by which the conversion process is expressed as a sequential estimation problem of tracking the target spectrum based on the observed source spectrum. The converted MFCC vectors are sequentially evaluated using a weighted sum of the target training set used as grid-points.**

To improve the perceived quality of the synthesized signals, we use a post-processing block for enhancing the global variance. Objective and subjective evaluations show that the enhanced-GB method is comparable to classic GMM-based methods, in terms of quality, and comparable to their enhanced versions, in terms of individuality.

I. Introduction

Voice conversion systems aim to modify the perceived identity of a source speaker saying a sentence, to that of a given target speaker. This kind of transformation is useful for personalization of Text-To-Speech systems, voice restoration in case of vocal pathology, and also for entertainment purposes.

In order to estimate a conversion function from a source speaker to a target speaker, voice conversion methods use training sets of both speakers. Most training algorithms require parallel and aligned data sets, that is, prerecorded sentences of the source and target speakers saying the same text, usually aligned using Dynamic Time Warping (DTW). A commonly used method for voice conversion is based on a linear conversion function, evaluated by a Gaussian Mixture Model (GMM) [1], [2]. This method, and other GMM-based conversion methods proposed since, use a linear conversion which produces over-smoothed spectral envelopes leading to muffled synthesized speech [3]. To reduce the muffling effect, methods for Global Variance (GV) enhancement of the spectral features have been proposed, integrated in the training process [4], [5], or as a post-processing block [6].

In this paper we propose a new method for spectral conversion based on a Grid-Based (GB) approximation [7]. As opposed to previously proposed methods, the GB conversion approach is trained using parallel sentences but is based on soft correspondence between the source and target vectors, obtained by phonetic labeling of the training sentences without frame alignment, thus eliminating the need for DTW.

GMM-based conversion methods are mostly trained using an iterative algorithm called Expectation Maximization, which often results in over-fitting [8]. These methods cannot be trained properly using small data sets, and their training stage may last hours or even days (depending on the amount of training data and computing platform), until convergence is achieved. Our GB method, however, is easily trained within seconds, using data sets of all sizes since its training stage is non iterative and involves simple computations based on the Euclidean distance between the training vectors.

In this paper we present an overall scheme, Enhanced-GB (En-GB), consisting of GB conversion followed by GV enhancement. Objectively, En-GB achieves similar spectral distortion and GV values as the classical GMM-based methods do. Listening tests show that En-GB achieves comparable quality to the classical GMM-based methods (without enhancement), and comparable individuality to their enhanced versions. Thus, the main advantages of the proposed approach are in the fast training, ability to work with small data sets, and the avoidance of time alignment.

This paper is organized as follows. In Sec. II, a brief description of GB approximation is presented. The new GB conversion method is described in Sec. III. Experimental results, demonstrating the performance of the proposed conversion method, in comparison to several other examined methods, are presented in Sec. IV. The paper is concluded in Sec. V.

II. Grid-Based Formulation

A brief formulation of sequential estimation using Bayesian tracking is presented in Sec. II-A. In many practical cases, applying this formulation yields a high computational load, which is sometimes unfeasible. The GB method provides a discrete approximation for Bayesian tracking with much less computational complexity, as described in Sec. II-B.

A. Bayesian Tracking

Denote by $\mathbf{y}_t$ a hidden state vector, following a first order Markov dynamics, and $\mathbf{x}_t$ as an observed signal, depending

978-1-4799-5988-4/14/$31.00 ©2014 IEEE

on the hidden state and on an i.i.d. measurement noise,$\mathbf{v}_t$:

$$\mathbf{y}_t = f_t(\mathbf{y}_{t-1}, \mathbf{u}_t) \quad (1)$$
$$\mathbf{x}_t = h_t(\mathbf{y}_t, \mathbf{v}_t), \quad (2)$$

where f_t and h_t are not necessarily linear and $\mathbf{u}_t$ is an i.i.d. noise sequence. The Bayesian optimal estimate for the state vector $\mathbf{y}_t$ in terms of mean squared error, given sequential samples of the observed signal,$\mathbf{x}_{1:t} \triangleq \{\mathbf{x}_1, ..., \mathbf{x}_t\}$, is obtained by:

$$\hat{\mathbf{y}}_t = E[\mathbf{y}_t|\mathbf{x}_{1:t}] = \int p(\mathbf{y}_t|\mathbf{x}_{1:t})\, \mathbf{y}_t d\mathbf{y}_t. \quad (3)$$

Assuming that the initial probability of the state vector is known and equal to the prior probability $p(\mathbf{y}_0) = p(\mathbf{y}_0|\mathbf{x}_0)$, the posterior probability $p(\mathbf{y}_t|\mathbf{x}_{1:t})$ can be obtained recursively in two stages:

$$p(\mathbf{y}_t|\mathbf{x}_{1:t-1}) = \int p(\mathbf{y}_t|\mathbf{y}_{t-1})\, p(\mathbf{y}_{t-1}|\mathbf{x}_{1:t-1})\, d\mathbf{y}_{t-1}$$
$$p(\mathbf{y}_t|\mathbf{x}_{1:t}) = \frac{p(\mathbf{x}_t|\mathbf{y}_t)\, p(\mathbf{y}_t|\mathbf{x}_{1:t-1})}{p(\mathbf{x}_t|\mathbf{x}_{1:t-1})}, \quad (4)$$

where,

$$p(\mathbf{x}_t|\mathbf{x}_{1:t-1}) = \int p(\mathbf{x}_t|\mathbf{y}_t)\, p(\mathbf{y}_t|\mathbf{x}_{1:t-1})\, d\mathbf{y}_t. \quad (5)$$

The likelihood function $p(\mathbf{x}_t|\mathbf{y}_t)$ is determined according to the function h_t and the statistics of the measurement noise $\mathbf{v}_t$.

When the noise signals $\mathbf{u}_t$ and $\mathbf{v}_t$ are Gaussian, and the functions $f_t(\cdot)$ and $h_t(\cdot)$ are linear and time invariant (meaning that $f_t(\cdot) \equiv f(\cdot)$ and $h_t(\cdot) \equiv h(\cdot)$), this recursion can be computed analytically, leading to Kalman filtering. Yet, in most practical cases where these conditions are not sustained, this derivation is hard and often performed using approximation methods such as GB approximation or particle filtering [7]. These methods sequentially evaluate the posterior probability as a discrete weighted sum using a given set of samples in case of GB, or a randomly drawn set in case of Particle Filtering.

B. Grid-Based Approximation

The main principle of GB approximation is to provide a Bayesian sequential estimation framework while avoiding the integral computations in (4) by using a discrete evaluation of the posterior probability.

Let $\left\{\mathbf{y}_t^k\right\}_{k=1}^{N_y}$ be a set of predetermined grid-points taken from the state-space $\left\{\mathbf{y}_t\right\}$. We divide the state space into cells, so that each cell has a grid point $\mathbf{y}_t^k$ as its center. Thus, the posterior probability can be approximated by:

$$p(\mathbf{y}_t|\mathbf{x}_{1:t}) \approx \sum_{k=1}^{N_y} w_{t|t}^k \delta\left(\mathbf{y}_t - \mathbf{y}_t^k\right). \quad (6)$$

where the posterior weights $\left\{w_{t|t}^k\right\}_{k=1}^{N_y}$ denote the conditional probabilities:

$$w_{t|t}^k = p\left(\mathbf{y}_t = \mathbf{y}_t^k|\mathbf{x}_{1:t}\right). \quad (7)$$

Using this discrete approximation, the prior probability is also approximated as a discrete sum:

$$p(\mathbf{y}_t|\mathbf{x}_{1:t-1}) \approx \sum_{k=1}^{N_y} w_{t|t-1}^k \delta\left(\mathbf{y}_t - \mathbf{y}_t^k\right). \quad (8)$$

The prior weights can be estimated sequentially [7]:

$$w_{t|t-1}^k \approx \sum_{l=1}^{N_y} w_{t-1|t-1}^l p\left(\mathbf{y}_t^k|\mathbf{y}_{t-1}^l\right), \quad (9)$$

where $p\left(\mathbf{y}_t^k|\mathbf{y}_{t-1}^l\right)$, called the *evidence probability*, is derived from the state space dynamics (eqn. (2)). The posterior weights $\{w_{t|t}^k\}_{k=1}^{N_y}$ are evaluated by:

$$w_{t|t}^k \approx \frac{w_{t|t-1}^k p\left(\mathbf{x}_t|\mathbf{y}_t^k\right)}{\sum_{l=1}^{N_y} w_{t|t-1}^l p\left(\mathbf{x}_t|\mathbf{y}_t^l\right)}, \quad (10)$$

where, as stated above, the likelihood probability $p\left(\mathbf{x}_t|\mathbf{y}_t^k\right)$ is derived from the measurement model (eqn. (1)).

Finally, the hidden state vector $\mathbf{y}_t$ is approximated using the posterior weights:

$$\hat{\mathbf{y}}_t = E[\mathbf{y}_t|\mathbf{x}_{1:t}] \approx \sum_{k=1}^{N_y} w_{t|t}^k \mathbf{y}_t^k. \quad (11)$$

III. Voice Conversion Using Grid-Based Approximation

We now use the GB approximation method described above as a framework for spectral voice conversion. We express the conversion as a sequential estimation problem, where the observed process is the source spectrum, and the tracked state-space is the target spectrum. We propose models for both likelihood and evidence densities, required for the sequential estimation process, as described in equations (9)-(11).

A. Training Stage

The training process described here includes pre-computation of the evidence and discrete likelihood probabilities, and is performed separately for every phoneme j, where $j = 1, ..., J$, and J is the overall number of phonemes. The source and target training sentences are assumed to be parallel and phonetically labeled. The spectral features of the two speakers are extracted from the voiced frames, but, as stated above, no time alignment is performed. Instead, a matching process of the source and target utterances is performed as follows. Each utterance r of a certain phoneme j at the source, is matched to its corresponding utterance at the target, according to the phonetic labeling. We avoid the transient nature of the beginning and ending of each utterance by using one third of the training vectors included in each utterance, extracted from the middle part. Based on these matched mid-utterances, we model the *discrete likelihood*

probability of a matched mid utterance r of phoneme j, used in eqn. (10), as:

$$p\left(\mathbf{x}_t=\mathbf{x}^m|\mathbf{y}_t=\mathbf{y}^k;j\right)=\begin{cases}\frac{1}{c_r^j} & \mathbf{x}^m,\mathbf{y}^k \text{ belong to the same mid-utterance r}\\ 0 & \text{otherwise}\end{cases} \tag{12}$$

where $\{\mathbf{x}^m;j\}_{m=1}^{N_x^j}$ and $\{\mathbf{y}^k;j\}_{k=1}^{N_y^j}$ are source and target training vectors, respectively, belonging to phoneme j, and c_r^j is the number of vectors related to utterance r at the target (i.e. $\sum_r c_r^j = N_y^j$). This definition ensures that the obtained discrete likelihood probability is normalized, i.e.:

$$\sum_{m=1}^{N_x^j} p\left(\mathbf{x}_t=\mathbf{x}^m|\mathbf{y}_t=\mathbf{y}^k;j\right)=1 \quad \forall k=1,...,N_y^j,\quad j=1,...,J \tag{13}$$

The discrete likelihood probability defines a relaxed correspondence between source and target training vectors, as opposed to a one-to-one match defined in other parallel methods, for which $p\left(\mathbf{x}_t=\mathbf{x}^m|\mathbf{y}_t=\mathbf{y}^k;j\right)=\delta_{m,k}$.

The evidence probability, as mentioned before, expresses the transition probability from state $\mathbf{y}^l$ to state $\mathbf{y}^k$. In natural speech, spectral feature vectors related to consecutive time frames are typically similar, but not identical. Motivated by this behavior, we model the transition probability as having the same value for all the states inside a ball, centered at $\mathbf{y}^k$ with a radius R_y. The probability of transitions to farther states, however, is taken as a simple Gaussian distribution, centered at $\mathbf{y}^k$. Altogether, we model the *discrete evidence probability*, used in eqn. (9), as:

$$\begin{aligned} p\left(\mathbf{y}_t=\mathbf{y}^k|\mathbf{y}_{t-1}=\mathbf{y}^l;j\right) &= \frac{1}{C_{evid}^{k,j}}e^{-\frac{M_{k,l}^2}{2}} \\ C_{evid}^{k,j} &\triangleq \sum_{k=1}^{N_y^j} e^{-\frac{M_{k,l}^2}{2}}, \end{aligned} \tag{14}$$

where j is the phoneme index; $k,l=1,...,N_y^j$, and where the exponential term in eqn. (14) is the maximum between the Mel Cepstral Distortion (MCD) of the two states $\mathbf{y}^l$ and $\mathbf{y}^k$ normalized by a parameter R_y, and 1:

$$M_{k,l} \triangleq \max\left(\frac{\text{MCD}\left(\mathbf{y}^k,\mathbf{y}^l\right)}{R_y},1\right), \tag{15}$$

$$\text{MCD}\left(\mathbf{y}^k,\mathbf{y}^l\right) \triangleq \frac{10\sqrt{2}}{\ln 10}\sqrt{\sum_{p=1}^{P}\left(y^k\left(p\right)-y^l\left(p\right)\right)^2}, \tag{16}$$

where $y^p\left(p\right)$ and $y^l\left(p\right)$ are the p-th elements of $\mathbf{y}^k$ and $\mathbf{y}^l$, respectively. An alternative approach would be to take the exponential term, defined in eqn. (15), as a normalized distance. For example, $M_{k,l}=\text{MCD}\left(\mathbf{y}^k,\mathbf{y}^l\right)/R_y$, where R_y is a parameter selected by the user. However, in case of a sparse training set the most substantial probability would be for staying in the same state. Since the training set is fixed, the likelihood and evidence densities are in fact time invariant.

B. Conversion Stage

The likelihood probability modeled above in eqn. (12) is defined only for a discrete set consisting of the source training vectors. In this section we extend (12) to model any input vector $\mathbf{x}_t\in\mathbb{R}^P$, as required by the GB formulation.

We model the continuous likelihood probability $p\left(\mathbf{x}_t|\mathbf{y}_t=\mathbf{y}^k;j\right)$ as a sum of the discrete likelihood probabilities $p\left(\mathbf{x}^m|\mathbf{y}_t=\mathbf{y}^k;j\right)$, $m=1,...,N_x^j$, (defined in (12) and (13)), each weighted by a Gaussian kernel, centered at $\mathbf{x}^m$:

$$\begin{aligned} &p\left(\mathbf{x}_t|\mathbf{y}_t=\mathbf{y}^k;j\right)= \\ &\quad\frac{1}{C_{LL}^{t,j}}\sum_{m=1}^{N_x^j}p\left(\mathbf{x}^m|\mathbf{y}_t=\mathbf{y}^k;j\right)e^{-\text{MCD}^2(\mathbf{x}_t,\mathbf{x}^m)/2R_x^2} \\ &C_{LL}^{t,j}\triangleq\sum_{k=1}^{N_y^j}p\left(\mathbf{x}_t|\mathbf{y}_t=\mathbf{y}^k;j\right), \end{aligned} \tag{17}$$

where R_x is a parameter determined by the user. The Gaussian term $e^{-\text{MCD}^2(\mathbf{x}_t,\mathbf{x}^m)/2R_x^2}$ can be viewed as an interpolation factor from the discrete space represented by the source training vectors to the continuous space of the test source vectors.

Define $w_{t|t}^{j,k}$ as the posterior weights corresponding to the training vectors $\{\mathbf{y}^k;j\}_{k=1}^{N_y^j}$, related to phoneme j:

$$w_{t|t}^{j,k}\triangleq p\left(\mathbf{y}_t|\mathbf{x}_{1:t};j\right). \tag{18}$$

During conversion, the posterior weights are sequentially evaluated, using the corresponding evidence and likelihood probabilities defined in (14) and (17), according to equations (9) and (10). The posterior weights are used to obtain the converted outcome as a discrete Bayesian approximation (as defined in (11)):

$$\mathcal{F}\{\mathbf{x}_t^j\}=E\left[\mathbf{y}_t|\mathbf{x}_{1:t};j\right]\approx\sum_{k=1}^{N_y^j}w_{t|t}^{j,k}\mathbf{y}_t^k. \tag{19}$$

where $\mathbf{x}_t^j$ belongs to a sequence of spectral features related to a certain test utterance of phoneme j. Due to the sequential update of the posterior weights, the converted spectral outputs evolve smoothly in time, within each utterance of a specific phoneme.

The main steps for converting a sequence of source vectors that belongs to phoneme j are summarized in Table I.

IV. Experimental Results

A. Experimental Conditions

In our experiments we used speech sentences of four U.S. English speakers taken from the CMU ARCTIC database [9]: two males (bdl, rms) and two females (clb, slt). The training

TABLE I
VOICE CONVERSION USING GB APPROXIMATION.

Input: a sequence of feature vectors

Initialization: set the initial weights, $\{w_{0|0}^{k}\}_{k=1}^{N_y^j}$

Main Iteration: for $t = 1, ...T$, perform the following steps:

1. Evaluate the prior weights, $\{w_{t|t-1}^{j,k}\}_{k=1}^{N_y^j}$, using equations (9) and (14)

2. Evaluate the posterior weights, $\{w_{t|t}^{j,k}\}_{k=1}^{N_y^j}$, using equations (10) and (12)

3. Evaluate $\tilde{\mathbf{y}}_t = \mathcal{F}\{\mathbf{x}_t^j\}$, using (19)

Output: a sequence of converted vectors - $\tilde{\mathbf{y}}_{1:T}$

set consists of 100 parallel sentences and the testing set consisted of 50 additional parallel sentences, all sampled at 16kHz and were phonetically labeled. Analysis, extraction of 25 Mel Frequency Cepstrum Coefficients (MFCCs) and synthesis were carried out using an available vocoder[1] based on the Harmonic Plus Noise model [10]. The sequences of the training data set used for GB conversion were matched (without alignment), as described in Sec. III-A. The training set used for the other examined methods, and the testing set, were each time aligned using a DTW algorithm based on phonetic labeling. Pitch was converted by a simple linear function using the mean and standard deviation values of the source and target speakers. Four conversion methods were examined: classical GMM-based conversion using joint training [2] (JGMM), classical GMM-based conversion using LS [1] (LS-GMM), Constrained GMM (CGMM) [5] and the GB conversion method proposed here.

B. Objective Evaluations

To obtain a fair comparison between different source-target pairs we normalized the mean spectral distortion between the converted and target signals by the mean spectral distortion between the source and target signals:

$$\mathrm{ND}\left(\tilde{\mathbf{Y}}_{1:T}, \mathbf{Y}_{1:T}\right) \triangleq \frac{\sum_{t=1}^{T} \mathrm{MCD}\left(\tilde{\mathbf{y}}_t, \mathbf{y}_t\right)}{\sum_{t=1}^{T} \mathrm{MCD}\left(\mathbf{x}_t, \mathbf{y}_t\right)}, \qquad (20)$$

where MCD is the distance between two cepstral vectors (defined in Sec. III, eqn. (16)) and $\tilde{\mathbf{Y}}_{1:T} \triangleq (\tilde{\mathbf{y}}_1, \tilde{\mathbf{y}}_2, \ldots, \tilde{\mathbf{y}}_T)^{\top}$, $\mathbf{Y}_{1:T} \triangleq (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T)^{\top}$ and $\mathbf{X}_{1:T} \triangleq (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)^{\top}$ are time aligned sequences of cepstral vectors, related to the converted, target, and source utterances, respectively. We use a Normalized Global Variance (NGV) to measure the variability of a sequence of converted vectors [6]:

$$\mathrm{NGV}\left\{\tilde{\mathbf{Y}}_{1:T}\right\} \triangleq \frac{1}{P} \sum_{p=1}^{P} \frac{\sigma^2_{\tilde{\mathbf{Y}}_{1:T}}(p)}{\sigma^2_{\mathbf{Y}}(p)}, \qquad (21)$$

where the global variance of each element p of a converted sequence $\tilde{\mathbf{Y}}_{1:T}$ is normalized by the empirical GV of the p-th element of the target speaker, obtained from the target training vectors. The desired values for these measures are ND $\rightarrow 0$ and NGV $\rightarrow 1$, indicating that the converted outcome is close to the target signal in terms of spectral similarity and global variance.

The GMM-based methods (LS-GMM, JGMM and CGMM) were trained using diagonal covariance matrices and $8, 16, 32, 64, 128, 256, 512$ Gaussian mixtures. The number of mixtures for the GMM-based methods, and the parameters of the proposed GB method (R_x and R_y) were selected for each method and training set so that a minimal ND was attained.

To further improve the quality of the synthesized speech, we applied a post-processing method for GV enhancement [6]. This method maximizes the GV of an input sequence, under a spectral distortion constraint. The GV of each enhanced sequence is increased up to the level where the MCD between the converted sequence and its enhanced version reaches a preset threshold value, denoted as θ_{MCD}. In this work we applied this GV enhancement method to LS-GMM, JGMM and to our proposed GB (tuned to minimal ND) conversion outcomes. The output signals of CGMM were not enhanced since NGV is already constrained to 1, in the training stage of this method.

Table II summarizes the main ND and NGV values achieved by the examined conversion methods, averaged over all four gender conversions: male-to-male (M2M), male-to-female (M2F), female-to-male (F2M) and female-to-female (F2F). The GB conversion followed by GV enhancement with $\theta_{MCD} = 1$dB (En-GB), produces similar NGV values to those attained by LS-GMM and JGMM (without enhancement), with slightly higher ND.

TABLE II
Objective performance: ND and NGV values using 100 training sentences, averaged over all four gender conversions (50 test sentences per gender conversion).

Conversion Method	ND	NGV
JGMM [2]	0.63	0.47
Enhanced JGMM	0.65	0.6
LS-GMM [1]	0.63	0.41
Enhanced LS-GMM	0.64	0.52
CGMM [5]	0.65	1.0
GB	0.67	0.35
Enhanced GB (En-GB)	0.69	0.44

The average training and conversion times of the examined methods are presented in Table III (using Matlab® software running on a Unix server with 48GB memory size and 2.5GHZ clock time). GMM-based methods are trained (128 mixtures) using an iterative method - Expectation Maximization (EM) - which lasts several hours till convergence is achieved. Note that in addition to EM, CGMM training also involves a significant computational cost due to a generalized SVD operation, required in the optimization process. The simplicity of GB's training stage, compared to the GMM-based methods, is well demonstrated as it lasts just seconds. The conversion times of all the examined methods, as well as the GV enhancement process, are very fast and last 23 msecs or less for a single sentence. Altogether, considering both training and conversion

[1] Available at http://aholab.ehu.es/ahocoder/

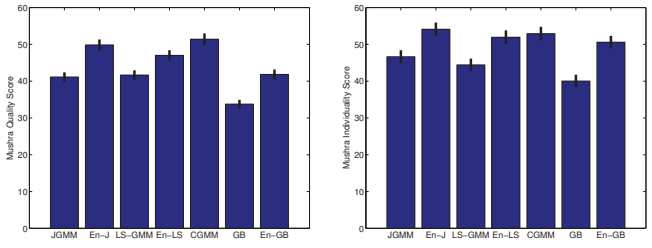
times, the proposed En-GB scheme is considerably faster than any of the other examined methods.

TABLE III
Average training times for 100 training sentences, and conversion times per frame, using Matlab® software running on a Unix server.

Method	Train. time	Conv. time per frame
JGMM [2]	7 h.	11 msec
LS-GMM [1]	8.5 h.	11 msec
CGMM [5]	11 h	11 msec
GB	**10 sec**	10 msec
GV enhancement [6]	**none in training**	23 msec

C. Subjective Evaluations

Listening tests were carried out to subjectively assess the performance of the examined methods (all trained by 100 sentences). The number of mixtures for the GMM-based methods, selected from among $8, 16, 32, 64, 128, 256, 512$, was set to 128 - for which the lowest ND was achieved. The proposed GB method was also tuned to minimal spectral distortion. We used informal listening tests to select the threshold value for GV enhancement from $\theta_{MCD} = 0.5, 1, 2, 4$dB. The best perceived quality was obtained with $\theta_{MCD} = 1$dB, for all the examined methods. We conducted subjective quality and individuality evaluations in a format similar to Multi Stimulus test with Hidden Reference and Anchor (MUSHRA) [11], as conducted by Godoy et. al. [12]. The listeners were presented with eight test signals: (a) a hidden reference - the target speaker; (b) JGMM; (c) Enhanced JMM; (d) LS-GMM; (e) Enhanced LS-GMM; (f) CGMM; (g) GB conversion; (h) Enhanced GB (En-GB). The test signals were randomly ordered, and the listeners were not informed about the hidden reference signals being included in the test set. During evaluation, the listeners were asked to compare the test signals to the reference signal (the target speaker) and rate their quality/similarity to the reference signal, between 0 to 100, where at least one of the test signals (the hidden reference) must be rated 100. As expected, all the listeners rated the hidden reference as 100. The mean scores of the examined methods averaged over all four gender conversions are presented in Figures 1(a) and 1(b), with their 95% confidence intervals. Without GV enhancement, LS-GMM and JGMM achieved higher quality scores than the proposed GB. Applying GV enhancement as a post processing block improved the score of all methods by 8%, on average. Still, CGMM was rated as having the best quality. Our overall conversion scheme, Enhanced GB, was rated as comparable to the classical GMM methods (without enhancement). Applying GV enhancement improved the individuality performance of JGMM and LS-GMM by 7.5% and the performance of GB by 11%. Altogether, the proposed En-GB method was marked as comparable to CGMM and to the enhanced versions of the classical GMM conversions.



(a) (b)

Fig. 1. Subjective tests averaged over all four gender conversions, comparing: JGMM [2], Enhanced JGMM (En-J), LS-GMM [1], Enhanced LS-GMM (En-LS), CGMM [5], GB and Enhanced GB (En-GB). (a) - quality test; (b) - individuality test.

V. Conclusion

We propose here a new method for spectral conversion, based on sequential Bayesian tracking, using a Grid-Based (GB) formulation. The target spectral evolution is modeled as a hidden Markov process, tracked by using the source spectrum, modeled as the observed process. As opposed to GMM-based methods, which are typically trained for hours or days, training GB is very simple and lasts just seconds; it does not require convergence of an iterative computation, and it is easily performed for both small and large scale databases. Additionally, although GB is trained using a parallel set, time alignment is not needed. We compared the proposed Enhanced GB scheme (GB followed by GV enhancement block) to CGMM and to classical GMM-based conversions, with and without GV enhancement. We showed that En-GB achieves comparable quality to the classical GMM-based methods (without enhancement), and comparable individuality to their enhanced versions.

References

[1] Y. Stylianou et al., "Continuous probabilistic transform for voice conversion," *IEEE Trans. on Speech and Audio Proc.*, vol. 6, no. 2, pp. 131–142, 1998.

[2] A. Kain and M. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. ICASSP*, 1998, pp. 285–288.

[3] T. Toda et al., "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," in *Proc. ICASSP*, 2005, pp. 9–12.

[4] T. Toda et al., "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 15, no. 8, pp. 2222–2235, 2007.

[5] H. Benisty and D. Malah, "Voice conversion using GMM with enhanced global variance," in *Proc. Interspeech*, 2011, pp. 669–672.

[6] H. Benisty et al., "Modular global variance enhancement for voice conversion systems," in *Proc. EUSIPCO*, 2012, pp. 370–374.

[7] M. S. Arulampalam et al., "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Transactions on Signal Proc.*, vol. 50, no. 2, pp. 174–188, 2002.

[8] L. Mesbashi et al., "Comparing GMM-based speech transformation systems," in *Proc. Interspeech*, 2007, pp. 1456–1989.

[9] J. Kominek and A. W. Black, "CMU ARCTIC databases for speech synthesis," 2003.

[10] O. Cappe and E. Moulines, "Regularization techniques for discrete cepstrum estimation," *IEEE Signal Processing Letters*, vol. 3, no. 4, pp. 100–102, 1996.

[11] "Multi stimulus test with hidden reference and anchors (MUSHRA)," Tech. Rep. ITU-R BS.1534-1, International Telecommunications Union, Jan. 2003.

[12] E. Godoy et al., "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 20, no. 4, pp. 1313–1323, 2012.