IRWIN AND JOAN JACOBS
CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGIES

Sequential Voice Conversion Using Grid-Based Approximation

Hadas Benisty, David Malah, and Koby Crammer

CCIT Report #843
November 2013

DEPARTMENT OF ELECTRICAL ENGINEERING
TECHNION - ISRAEL INSTITUTE OF TECHNOLOGY, HAIFA 32000, ISRAEL

CCIT Report #843 November 2013

Sequential Voice Conversion Using Grid-Based Approximation

Hadas Benisty, David Malah, and Koby Crammer

Abstract

The goal of voice conversion is to modify a source speaker's speech to sound as if spoken by a target speaker. Common conversion methods are based on Gaussian Mixture Modeling (GMM), which require exhaustive training (typically lasting hours), often leading to ill-conditioning, if the dataset used is too small. Additionally, the training process is based on a one-to-one match between the source and target vectors, requiring time alignment. We propose a new conversion method that is trained in seconds, using either small or large scale datasets (50-200 sentences). It requires a parallel dataset but without time alignment. 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.

1 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 (TTS) systems, voice restoration in case of vocal pathology, obtaining a false identity when answering the phone (for safety reasons, for example), and also for entertainment purposes such as online role-playing games.

The identity of a speaker is associated with the spectral envelope of the speech signal, and with its prosody attributes: pitch, duration, and energy. Most voice conversion methods aim to transform the spectral envelope of the source speaker to the spectral envelope of the target speaker. The pitch contour is commonly converted by a linear transformation based on the global mean and standard deviation values of the pitch frequency.

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 data sets, that is, prerecorded sentences of the source and target speakers saying the same text. In such a setup, evaluation of a conversion function is based on coupled feature vectors - source and target. However, since the two speakers generally do not pronounce the text at the exact same rate, matching an analysis frame of the source speaker to one of the analysis frames of the target speaker is not straightforward. A time alignment is usually carried out using Dynamic Time Warping (DTW), constrained by starting and ending of speech utterances [12]. These time stamps are commonly obtained by phonetic labeling, representing the beginning and ending of each phoneme. When phonetic labeling is unavailable, Voice Activity Detection (VAD) is applied so the time stamps generated by the VAD represent the beginning and ending of each word. Since the source and target training sentences are not spoken in exactly the same rate, DTW often replicates or omits feature vectors, artificially producing a match. The importance of correct time alignment was recently demonstrated as having a large influence on the quality of the synthesized converted speech [15]. A different approach was suggested by [24], where a statistical model for an eigen-voice was trained using several parallel data-sets. The conversion function is trained using the eigen-voice model and speech sentences related to a target speaker (not necessarily parallel to the source data-sets).

One of the earlier approaches for spectral conversion uses a codebook representation of the spectral features obtained from a parallel training set [2]. Due to the limited codebook size, the converted spectral envelope is deficiently represented which leads to poor quality synthesized speech. Later, a more flexible approach for spectral conversion, based on a Gaussian Mixture Model (GMM), was proposed [21] and is the most commonly used method to date. The source training data is used to train a GMM, and the linear conversion function is evaluated by Least Squares (LS) using a parallel and time-aligned training set. Alternatively, these conversion parameters may be evaluated using a joint source-target GMM training [16]. These linear conversion methods produce over-smoothed spectral envelopes leading to muffled synthesized speech ([22],[25]). Several modifications of the GMM-based conversion have been proposed since, among these: GMM with Dynamic Frequency Warping (DFW) [25], GMM and codebook selection [17] and a combined pitch and spectral envelope GMM-based conversion [9]. Still, these GMM-based conversion methods have been reported to produce muffled output signals, probably due to excessive smoothing of the temporal evolution of the spectral envelope. Recently, a different approach aiming to capture the temporal evolution of the spectral envelope was presented [23]. A GMM is trained using concatenated sequences of the source and target spectral features, and the conversion function is evaluated using Maximum

Likelihood (ML) estimation. To reduce the muffling effect, the Global Variance (GV) of the spectral features was considered in the trained statistical model. A GV enhancement method was also proposed in the framework of the classical GMM-based conversion, where the GV of the converted features is constrained to match the GV of the features related to the target speaker [5]. These two conversion schemes (with integrated GV enhancement) improve the quality of the converted signals, at the expense of some increase in the spectral distance between the converted and target signals.

In this paper we propose a new method for spectral conversion based on a Grid-Based (GB) approximation [4]. We express the spectral conversion process as a sequential Bayesian estimation problem of tracking the target spectrum using observed samples from the source spectrum. We propose models for evaluation of the evidence and likelihood probabilities needed for the GB formulation. Using these approximated probabilities the algorithm sequentially evaluates the converted spectrum as a weighted sum of the target training vectors.

As opposed to previously proposed methods that use parallel and time aligned training sets, the GB conversion approach does not require a one-to-one correspondence between the source and target training vectors. The training process uses 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 [8], which often results in overfitting [20]. 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.

The GB conversion proposed here provides various working points in terms of spectral distortion and GV. When tuned to minimal spectral distortion, GB achieves comparable performance to the classical GMM-based methods ([21], [16]), in terms of spectral distortion and GV. When tuned to maximal GV, GB produces higher values of GV (0.8 of its natural value for the target speaker), at the expense of increased spectral distortion. Informal listening tests showed that higher synthesis quality is obtained when GB is tuned to minimal spectral distortion. To further improve the quality we applied a GV enhancement post-processing block. We recently proposed this GV enhancement approach and examined its effect on signals converted by a classical GMM conversion method [6]. In this paper we present an overall scheme, Enhanced-GB (En-GB), consisting of GB conversion (tuned

for minimal spectral distortion), followed by GV enhancement. We used objective measures and also performed extensive subjective evaluations, to compare the proposed En-GB conversion method to several GMM-based conversion methods, with and without enhancement. Objectively, En-GB achieves similar spectral distortion and GV values as the classical GMM-based methods do (without enhancement). Listening tests show that in terms of quality, En-GB is comparable to the classical GMM-based conversion methods (without enhancement). Furthermore, in terms of similarity to the target speaker, the En-GB scheme is comparable to the enhanced versions of the classical GMM-based methods. Thus, the main advantages of the proposed approach are in the short training, ability to work with small data sets, and the avoidance of time alignment of frames in parallel data.

This paper is organized as follows. In Sec. 2, a brief description of GB approximation is presented. The new GB conversion method is described in Sec. 3. Experimental results, demonstrating the performance of the proposed conversion method and the effect of GV enhancement on its converted output signals, in comparison to several other examined methods, are presented in Sec. 4. Conclusions and further research suggestions are given in Sec. 5.

2 Grid-Based Formulation

A brief formulation of sequential estimation using Bayesian tracking is presented in Sec. 2.1. 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. 2.2.

2.1 Bayesian Tracking

Denote by $\mathbf{y}_t$ a hidden state vector, following a first order Markov dynamics:

$$\mathbf{y}_t = f_t\left(\mathbf{y}_{t-1}, \mathbf{u}_t\right), \tag{1}$$

where f_t is a function (not necessarily linear) of $\mathbf{y}_{t-1}$ and of an i.i.d. noise sequence $\mathbf{u}_t$. The observed signal, $\mathbf{x}_t$, depends on the hidden state and on an i.i.d. measurement noise,$\mathbf{v}_t$:

$$\mathbf{x}_t = h_t\left(\mathbf{y}_t, \mathbf{v}_t\right), \tag{2}$$

where $h_t\left(\cdot\right)$ may also be non-linear.

Denote by $\mathbf{x}_{1:t}$ as t vectors sequentially sampled from the observed process - $\mathbf{x}_{1:t} \triangleq \{\mathbf{x}_1, ..., \mathbf{x}_t\}$. Assuming that the initial probability of the state vector, $p\left(\mathbf{y}_0\right)$, is known and equal to the prior probability $p\left(\mathbf{y}_0\right) = p\left(\mathbf{y}_0|\mathbf{x}_0\right)$, the posterior probability $p\left(\mathbf{y}_t|\mathbf{x}_{1:t}\right)$ can be obtained recursively in two stages:

1. Prediction - obtain the prior probability:

$$p\left(\mathbf{y}_t|\mathbf{x}_{1:t-1}\right) = \int p\left(\mathbf{y}_t|\mathbf{y}_{t-1}\right) p\left(\mathbf{y}_{t-1}|\mathbf{x}_{1:t-1}\right) d\mathbf{y}_{t-1}. \tag{3}$$

2. Update - use the current observation $\mathbf{x}_t$ to update the posterior probability:

$$p\left(\mathbf{y}_t|\mathbf{x}_{1:t}\right) = \frac{p\left(\mathbf{x}_t|\mathbf{y}_t\right) p\left(\mathbf{y}_t|\mathbf{x}_{1:t-1}\right)}{p\left(\mathbf{x}_t|\mathbf{x}_{1:t-1}\right)}, \tag{4}$$

where,

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

The likelihood function $p\left(\mathbf{x}_t|\mathbf{y}_t\right)$ is determined according to the measurement model (eqn. (2)) and the statistics of the measurement noise $\mathbf{v}_t$. The Bayesian optimal estimate for the state vector $\mathbf{y}_t$ in terms of mean squared error is obtained by[1]:

$$\hat{\mathbf{y}}_t = E\left[\mathbf{y}_t|\mathbf{x}_{1:t}\right] = \int p\left(\mathbf{y}_t|\mathbf{x}_{1:t}\right) \mathbf{y}_t d\mathbf{y}_t. \tag{6}$$

When the noise signals $\mathbf{u}_t$ and $\mathbf{v}_t$ are Gaussian, and the functions $f_t\left(\cdot\right)$ and $h_t\left(\cdot\right)$ are linear and time invariant (meaning that $f_t\left(\cdot\right) \equiv f\left(\cdot\right)$ and $h_t\left(\cdot\right) \equiv h\left(\cdot\right)$), this recursion can be computed analytically, leading to Kalman filtering [3]. 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 [4]. 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.

In this paper, we express the spectral conversion process as a sequential estimation problem tracking the target spectrum, using observed samples from the source spectrum. We propose models for the evidence and likelihood probabilities needed for the GB formulation. Using these approximated probabilities the algorithm sequentially evaluates the converted spectrum as a weighted sum of the target training vectors. It is well known that the performance of particle filtering crucially depends on successful statistical modeling of the state-space temporal evolution. The performance of GB, on the other hand, depends on dense modeling of the state-space by a set of predetermined grid-points. Since a diverse training set is usually available in most conversion setups, we apply the GB approximation method, using the target training vectors as grid-points, as described below.

[1] In general, any arbitrary integrable function of the state vector $\mathbf{y}_t$ can be evaluated [4].

2.2 Grid-Based Approximation

The main principle of GB approximation is to provide a Bayesian sequential estimation framework while avoiding the integral computations in (3) and (5) 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[2]:

$$p\left(\mathbf{y}_t|\mathbf{x}_{1:t}\right) \approx \sum_{k=1}^{N_y} w_{t|t}^k \delta\left(\mathbf{y}_t - \mathbf{y}_t^k\right). \tag{7}$$

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

$$w_{t|t}^k = p\left(\mathbf{y}_t = \mathbf{y}_t^k|\mathbf{x}_{1:t}\right). \tag{8}$$

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

$$p\left(\mathbf{y}_t|\mathbf{x}_{1:t-1}\right) \approx \sum_{k=1}^{N_y} w_{t|t-1}^k \delta\left(\mathbf{y}_t - \mathbf{y}_t^k\right). \tag{9}$$

The prior weights can be estimated sequentially [4]:

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

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. (1)). The posterior weights $\{w_{t|t}^k\}_{k=1}^{N_y}$ are evaluated by:

$$w_{t|t}^k \approx \frac{w_{t|t-1}^k p\left(\mathbf{x}_t|\mathbf{y}_t^k\right)}{\sum_{l=1}^{N_y} w_{t|t-1}^l p\left(\mathbf{x}_t|\mathbf{y}_t^l\right)}, \tag{11}$$

where, as stated above, the likelihood probability $p\left(\mathbf{x}_t|\mathbf{y}_t^k\right)$ is derived from the measurement model (eqn. (2)).

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

$$\hat{\mathbf{y}}_t = E\left[\mathbf{y}_t|\mathbf{x}_{1:t}\right] \approx \sum_{k=1}^{N_y} w_{t|t}^k \mathbf{y}_t^k. \tag{12}$$

[2] If the state space is indeed discrete and finite, and the grid-points consist of all its states, this evaluation becomes exact.

Note that equations (10), (11) and (12) are discrete evaluations of equations (3)-(6), correspondingly. It is known [4] that the estimated terms in (7) and in (12) are biased for any finite N_y. Still, as more grid points are taken the bias gets smaller and the approximation improves, since the state space is more densely represented.

The sequential estimation process is initialized using the initial probability of the state vector $p\left(\mathbf{y}_0^k\right)$, which as stated above, is assumed to be known:

$$w_{0|0}^k = p\left(\mathbf{y}_0^k\right). \tag{13}$$

Table 1 summarizes the main stages of sequential Bayesian estimation using GB approximation.

Tab. 1: Bayesian Estimation Using Grid-Based Approximation.

Input: a sequence of states sampled from the observed process - $\mathbf{x}_{1:T}$
Initialization: set the initial weights, $\{w_{0|0}^k\}_{k=1}^{N_y}$, using eqn. (13)
Main Iteration: for $t = 1, ...T$, perform the following steps:
1. Evaluate the prior weights, $\{w_{t|t-1}^k\}_{k=1}^{N_y}$, using eqn. (10).
2. Evaluate the posterior weights, $\{w_{t|t}^k\}_{k=1}^{N_y}$, using eqn. (11).
3. Evaluate the hidden state, $\hat{\mathbf{y}}_\mathbf{t}$, using eqn. (12).
Output: a sequence of the estimated hidden states - $\hat{\mathbf{y}}_{1:T}$

3 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 (10)-(12). The GB conversion method proposed here uses a parallel training set, but does not require time alignment between the source and target training vectors since it is trained using soft correspondence between them, rather than matched pairs. The training and conversion stages of the proposed GB conversion method are presented below in Secs. 3.1 and 3.2, respectively.

3.1 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. (11), as:

$$p\left(\mathbf{x}_t = \mathbf{x}^m | \mathbf{y}_t = \mathbf{y}^k; j\right) = \begin{cases} \frac{1}{c_r^j} & \mathbf{x}^m, \mathbf{y}^k \text{ belong to the same mid-utterance r} \\ 0 & \text{otherwise,} \end{cases} \tag{14}$$

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.:

$$\sum_{m=1}^{N_x^j} p\left(\mathbf{x}_t = \mathbf{x}^m | \mathbf{y}_t = \mathbf{y}^k; j\right) = 1, \quad \forall k = 1, ..., N_y^j, \quad j = 1, ..., J. \tag{15}$$

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. (10), as:

$$\begin{aligned} p\left(\mathbf{y}_t = \mathbf{y}^k | \mathbf{y}_{t-1} = \mathbf{y}^l; j\right) &= \frac{1}{C_{evid}^{k,j}} e^{-\frac{M_{k,l}^2}{2}} \\ C_{evid}^{k,j} &\triangleq \sum_{k=1}^{N_y^j} e^{-\frac{M_{k,l}^2}{2}}, \end{aligned} \tag{16}$$

where j is the phoneme index; $k, l = 1, ..., N_y^j$, and where the exponential term in eqn. (16) 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:

$$M_{k,l} = \max\left(\frac{\mathrm{MCD}\left(\mathbf{y}^k,\mathbf{y}^l\right)}{R_y},1\right), \tag{17}$$

$$\mathrm{MCD}\left(\mathbf{y}^k,\mathbf{y}^l\right) = \frac{10\sqrt{2}}{\ln 10}\sqrt{\sum_{p=1}^{P}\left(y^k\left(p\right)-y^l\left(p\right)\right)^2}, \tag{18}$$

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. (17), as a normalized distance. For example, $M_{k,l}=\mathrm{MCD}\left(\mathbf{y}^k,\mathbf{y}^l\right)/R_y$, 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.

3.2 Conversion Stage

The likelihood probability modeled above in eqn. (14) is defined only for a discrete set consisting of the source training vector. In this section we extend (14) to model any input vector $\mathbf{x}_t \in \mathtt{R}^P$, 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 (14) and (15)), each weighted by a Gaussian kernel, centered at $\mathbf{x}^m$:

$$p\left(\mathbf{x}_t|\mathbf{y}_t=\mathbf{y}^k;j\right) = \frac{1}{C_{LL}^{t,j}}\sum_{m=1}^{N_x^j}p\left(\mathbf{x}^m|\mathbf{y}_t=\mathbf{y}^k;j\right)e^{-\mathrm{MCD}^2(\mathbf{x}_t,\mathbf{x}^m)/2R_x^2}$$

$$C_{LL}^{t,j} \triangleq \sum_{k=1}^{N_y^j}p\left(\mathbf{x}_t|\mathbf{y}_t=\mathbf{y}^k;j\right), \tag{19}$$

where R_x is a parameter determined by the user. The Gaussian term $e^{-\mathrm{MCD}^2(\mathbf{x}_t,\mathbf{x}^m)/2R_x^2}$ 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:

$$w_{t|t}^{j,k} \triangleq p\left(\mathbf{y}_t|\mathbf{x}_{1:t};j\right). \tag{20}$$

During conversion, the posterior weights are sequentially evaluated, using the corresponding evidence and likelihood probabilities defined in (16) and (19), according to equations (10) and (11). The posterior weights are used

to obtain the converted outcome as a discrete Bayesian approximation (as defined in (12)):

$$\mathcal{F}\{\mathbf{x}_t^j\} = E\left[\mathbf{y}_t|\mathbf{x}_{1:t}; j\right] \approx \sum_{k=1}^{N_y^j} w_{t|t}^{j,k}\mathbf{y}_t^k. \tag{21}$$

where $\mathbf{x}_t^j$ belongs to a sequence of spectral features related to a certain test utterance of phoneme j.

As mentioned above, the training set of each phoneme is composed of feature vectors extracted from the middle part of each utterance. Nevertheless, during conversion, all vectors in each utterance $\mathbf{x}_{1:T}^j$ are converted using these mid-utterance vectors as grid-points. Due to the sequential update of the posterior weights, the converted spectral outputs evolve smoothly in time, within each utterance of a specific phoneme.

Inter-Phoneme Evolution:

In order to maintain a smooth evolution also during the transition between consecutive phonemes, we evaluate the initial condition for each utterance of a certain phoneme according to the preceding phoneme. Denote by $\mathbf{x}_{1:T'}^i$ and $\mathbf{x}_{1:T}^j$ two consecutive sequences of source feature vectors representing an utterance of phoneme i followed by an utterance of phoneme j, respectively. If the preceding utterance is unvoiced or silence (i.e., not related to a voiced phoneme), a uniform initial condition is taken for the conversion of the current utterance: $w_{0|0}^{j,k} = 1/N_y^j$ for $k = 1, ..., N_y^j$. If the preceding utterance is a voiced phoneme, its feature vectors are converted using its corresponding set of grid-points $\{\mathbf{y}^k; i\}_{k=1}^{N_y^i}$. The current utterance, however, is converted according to the grid-points related to the phoneme j, $\{\mathbf{y}^k; j\}_{k=1}^{N_y^j}$. Therefore, the weights of the current utterance cannot be directly initialized using the weights of the preceding utterance. To resolve this problem, we re-evaluate the weights of the last vector of the preceding utterance, $\mathbf{x}_{T'}^i$, as if it was converted using the grid-points of the current phoneme j. The initial weights for converting the first vector of the current utterance, $\mathbf{x}_1^j$, are taken as the posterior probabilities of the vector $\mathbf{x}_{T'}^i$, as if it was converted using the phoneme j:

$$w_{0|0}^{j,k} = p\left(\mathbf{y}_{T'} = \mathbf{y}^k|\mathbf{x}_{T'}^i; j\right), \quad k = 1, ..., N_y^j, \tag{22}$$

where these probabilities are evaluated using uniform weights as an initial condition:

$$w_{-1|-1}^{j,k} \triangleq p\left(\mathbf{y}_{T'-1} = \mathbf{y}^k|\mathbf{x}_{T'-1}^i; j\right) = \frac{1}{N_y^j}. \tag{23}$$

Figure 1 demonstrates the obtained time evolution of the first and third

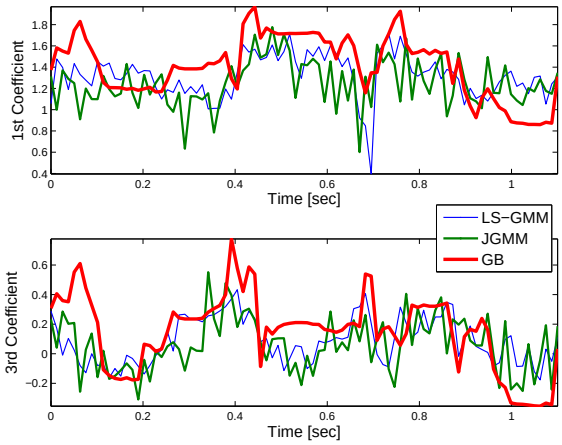

Fig. 1: Temporal evolution of the 1st and 3rd coefficients of the words converted by: LS-GMM - blue thin line; JGMM - green; GB - red thick line.

MFCCs using GB conversion, compared to the classical GMM-based conversions: LS-GMM [21] and JGMM [16]. The classical GMM-based conversions are applied frame by frame which may lead to discontinuities. The proposed GB, however, is based on a sequential update leading to a smoother time evolution of the cepstral elements, as seen in Fig. 1.

To conclude, the main stages of converting a sequence of source vectors that belongs to phoneme j are summarized in Table 2.

Tab. 2: Voice Conversion Using GB Approximation.

Input: a sequence of feature vectors related to the current phoneme and the last source vector related to the preceding phoneme, correspondingly: $\mathbf{x}^j_{1:T}, \mathbf{x}^i_{T'}$

Initialization: set the initial weights, $\{w^k_{0|0}\}_{k=1}^{N^j_y}$, according to (22) and (23).

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

1. Evaluate the prior weights, $\{w^{j,k}_{t|t-1}\}_{k=1}^{N^j_y}$, using equations (10) and (16).
2. Evaluate the posterior weights, $\{w^{j,k}_{t|t}\}_{k=1}^{N^j_y}$, using equations (11) and (14).
3. Evaluate $\tilde{\mathbf{y}}_t = \mathcal{F}\{\mathbf{x}^j_t\}$, using (21).

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

4 Experimental Results

4.1 Experimental Conditions

In our experiments we used speech sentences of four U.S. English speakers taken from the CMU ARCTIC database [18]: two males (bdl, rms) and

two females (clb, slt). Three different sizes of training sets: 50, 100 and 200 parallel sentences were used to demonstrate the performance of the examined methods as a function of training set size. The testing set consisted of 50 additional parallel sentences. All sentences were sampled at 16kHz and were phonetically labeled.

Analysis and synthesis were both carried out using an available vocoder [10]. This vocoder uses a two-band harmonic/noise parametrization, separated by a maximal voicing frequency for representing each spectral envelope [13]. 25 Mel Frequency Cepstrum Coefficients (MFCCs) were extracted from the harmonic parameters [7]: the zero-th coefficients, related to the energy, were not converted. The other 24 coefficients were used as spectral feature vectors during training and conversion.

The spectral features of unvoiced frames were not converted but simply copied to the converted sentence, since they do not capture much of the speaker's individuality [19] and their conversion often leads to quality degradation [11]. The maximal voicing frequency was also not converted but re-estimated from the converted parameters by the vocoder. The sequences of the training data set used for GB conversion were matched (without alignment), as described in Sec. 3.1. 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 [12].

Pitch was converted by a simple linear function using the mean and standard deviation values of the source and target speakers,

$$\hat{f}_0^{(y),t} = \mu^{(y)} + \left(\sigma^{(y)}/\sigma^{(x)}\right)\left(f_0^{(x),t} - \mu^{(x)}\right), \tag{24}$$

where $f_0^{(x),t}$ and $\hat{f}_0^{(y),t}$ are the pitch values of the source and converted signals at the t-th frame, respectively. The parameters $\mu^{(x)}$ and $\mu^{(y)}$ are the mean pitch values, and $\sigma^{(x)}$ and $\sigma^{(y)}$ are the standard deviations of the source and target pitch values, respectively. In this case the mean and standard deviation of the converted pitch contour match the mean and standard deviation of the pitch values of the target speaker.

Four conversion methods were examined: classical GMM-based conversion using joint training [16] (JGMM), classical GMM-based conversion using LS [21] (LS-GMM), Constrained GMM (CGMM) [5] and the GB conversion method proposed here.

4.2 Objective Evaluations

We evaluated the performance of the examined conversion methods by two objective measures: Normalized Distortion (ND) and Normalized GV (NGV), as defined below.

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 [26]:

$$\mathrm{ND}\left(\tilde{\mathbf{Y}}_{1:T}, \mathbf{Y}_{1:T}\right) = \frac{\sum_{t=1}^{T} \mathrm{MCD}\left(\tilde{\mathbf{y}}_t, \mathbf{y}_t\right)}{\sum_{t=1}^{T} \mathrm{MCD}\left(\mathbf{x}_t, \mathbf{y}_t\right)}, \tag{25}$$

where MCD is the distance between two cepstral vectors (defined in Sec. 3, eqn. (18)) 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.

The Global Variance (GV) of the p-th elements of a sequence, $\tilde{\mathbf{Y}}_{1:T}$, representing a converted speech utterance, is:

$$\sigma^2_{\tilde{\mathbf{Y}}_{1:T}}(p) = \frac{1}{T}\sum_{t=1}^{T}\left(\tilde{y}_t(p) - \frac{1}{T}\sum_{\tau=1}^{T}\tilde{y}_\tau(p)\right)^2, \tag{26}$$

In this paper we use a Normalized Global Variance (NGV) to measure the variability of a sequence of converted vectors:

$$\mathrm{NGV}\left\{\tilde{\mathbf{Y}}_{1:T}\right\} \triangleq \frac{1}{P}\sum_{p=1}^{P}\frac{\sigma^2_{\tilde{\mathbf{Y}}_{1:T}}(p)}{\sigma^2_{\mathbf{Y}}(p)}, \tag{27}$$

where $\sigma^2_{\mathbf{Y}}(p)$ is the empirical GV of the p-th elements of the target speaker, obtained from the target training vectors:

$$\sigma^2_{\mathbf{Y}}(p) = \frac{1}{N_y}\sum_{k=1}^{N_y}\left(y^k(p) - \frac{1}{N_y}\sum_{n=1}^{N_y}y^n(p)\right)^2. \tag{28}$$

Note that the target GV defined in eqn. (28) is evaluated by averaging over the entire training corpus. This evaluation of GV is different from a recently proposed approach [23] for spectral conversion and GV enhancement, where the GV of each utterance of the target is modeled as a random variable drawn from a Gaussian distribution.

The desired values for these measures are ND $\to$ 0 and NGV $\to$ 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 was selected for each method and training set so that a minimal ND was attained.

Figure 2 presents the ND vs. NGV values obtained for LS-GMM, JGMM, CGMM and the proposed GB, all trained using 100 sentences, for a male-to-male conversion. The classical GMM-based conversion methods, LS-GMM and JGMM, produce relatively low ND, but suffer from very low NGV. The

proposed GB conversion method provides a range of possible ND and NGV combinations: as the parameters R_x and R_y get smaller, less grid-points are considered in the weighted sum, so the NGV increases, but the ND also increases. In terms of the examined objective measures, CGMM outperforms all the examined methods since it produces higher NGV and lower ND at the same time.

The ND and NGV values attained by the examined methods, as a function of the size of the training set, are presented in Table 3. Training CGMM, as previously presented [5], involves a high computational load due to a generalized SVD operation required in the optimization process. Consequently, results for this methods are presented here only for 50 and 100 training sentences. The performance of the proposed GB conversion method is demonstrated using two extremal working points: one is maximal NGV (GB Max-NGV, attained for $R_x = 1$dB, $R_y = 2$dB) and the other is minimal ND (GB Min-ND, attained for $R_x = 4$dB, $R_y = 6$dB). Informal listening tests showed that the proposed GB method achieves higher quality when tuned to minimal spectral distortion than to maximal GV. Although the actual ND and NGV values achieved by each method, as indicated in Table 3 are very similar, still several trends can be observed: adding more training sentences improves the mean spectral similarity to the target for all the examined methods; the most significant improvement is achieved by the proposed GB. In terms of NGV, using over 100 training sentences slightly increases the NGV for CGMM and GB Max-NGV, yet for LS-GMM and JGMM the NGV decreases.

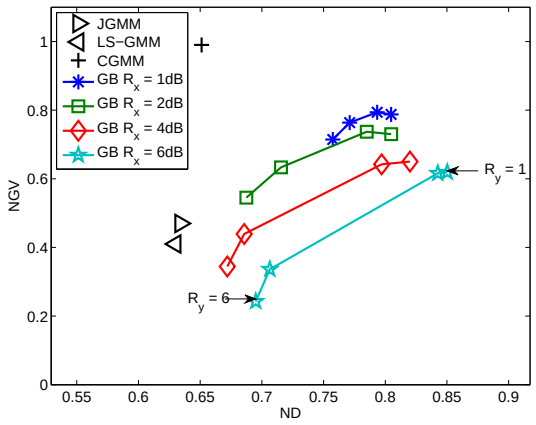

Fig. 2: ND vs. NGV using 100 training sentences for a male-to-male conversion: right triangle - JGMM [16]; left triangle - LS-GMM [21]; plus sign - CGMM [5]; asterisk, square, diamond and star - the proposed GB conversion using $R_x = (1, 2, 4, 6)$ [dB], correspondingly, for $R_y = (1, 2, 4, 6)$[dB].

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}. We recently showed [6] that this method leads to significant improvement in the perceived quality of signals converted by LS-GMM. 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 4 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, tuned to minimal ND and 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.

Tab. 3: *Objective performance: ND and NGV for male-to-male conversion using 50, 100 and 200 training sentences.*

	ND			NGV		
No. of Training Sentences	50	100	200	50	100	200
JGMM [16]	0.65	0.63	0.62	0.37	0.38	0.38
LS-GMM [21]	0.64	0.62	0.62	0.32	0.33	0.31
CGMM [5]	0.69	0.68	-	0.82	0.84	-
GB Min-ND	0.69	0.67	0.64	0.3	0.27	0.25
GB Max-NGV	0.82	0.80	0.78	0.66	0.65	0.69

Tab. 4: *Objective performance: ND and NGV values using 100 training sentences, averaged over all four gender conversions.*

Conversion Method	ND	NGV
JGMM [16]	0.63	0.47
Enhanced JGMM	0.65	0.6
LS-GMM [21]	0.63	0.41
Enhanced LS-GMM	0.64	0.52
CGMM [5]	0.65	1.0
GB Min-ND	0.67	0.35
Enhanced GB Min-ND En-GB	0.69	0.44

The average training and conversion times of the examined methods,

using 50,100 and 200 sentences, are presented in Table 5 (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 [8] (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 times, the proposed En-GB scheme is considerably faster than any of the other examined methods.

Tab. 5: *Average training times for 50, 100 and 200 training sentences, and conversion times per frame, using Matlab® software running on a Unix server.*

Method	Training time			Conversion time per frame
No. of Training Sentences	50	100	200	
JGMM [16]	3 h.	7 h.	8 h.	11 msec
LS-GMM [21]	2.5 h.	8.5 h.	10.5 h.	11 msec
CGMM [5]	3.5 h.	11 h.	-	11 msec
GB	**2 sec**	**10 sec**	**20 sec**	10 msec
GV enhancement [6]	**none in training**			23 msec

To conclude the objective examination, in terms of ND vs. NGV, the CGMM conversion method outperforms all the examined methods since it produces a higher NGV together with a lower ND at the same time. Nevertheless, we note that the proposed En-GB scheme achieves comparable objective performance to the classical GMM methods, while its training time is significantly shorter.

In the next section we present subjective evaluation results comparing the proposed En-GB conversion scheme to the classical GMM-based conversion methods (with and without enhancement) and to CGMM, in terms of perceived quality and similarity to the target speaker.

4.3 Subjective Evaluations

Listening tests were carried out to subjectively assess the performance of the examined methods (all trained by 100 sentences). The same four speakers (two males and two females) that were used for the objective evaluations, were used for the subjective evaluations. 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 (GB Min-ND). For simplicity of notation, GB will refer to this condition from this point on, unless otherwise stated. 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. All four gender conversions were performed using the same parameters values as described above.

We conducted subjective quality evaluations in a format similar to Multi Stimulus test with Hidden Reference and Anchor (MUSHRA) [1]. 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 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 for M2M, M2F, F2M and F2F conversions, and also their scores averaged over all four conversions are presented in Figures 3 and 4, respectively. All subjective results are presented with their 95% confidence intervals.

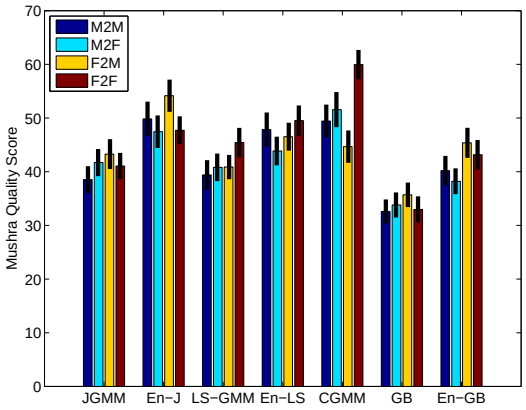

Fig. 3: Subjective quality test, comparing: JGMM [16], Enhanced JGMM (En-J), LS-GMM [21], Enhanced LS-GMM (En-LS), CGMM [5], GB and Enhanced GB (En-GB).

Without 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, En-

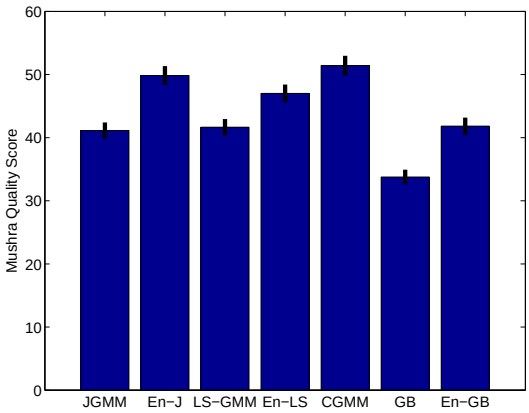

Fig. 4: Subjective quality test averaged over all four gender conversions comparing: JGMM [16], Enhanced JGMM (En-J), LS-GMM [21], Enhanced LS-GMM (En-LS), CGMM [5], GB and Enhanced GB (En-GB).

hanced GB, was rated as comparable to the classical GMM methods (without enhancement).

We evaluated the individuality performance using, again, a similar format to MUSHRA, as conducted by Godony et. al. [14]. The listeners were presented with the same test signals (including the hidden reference) and were asked to rate their similarity to the reference signal, in terms of the speaker's identity, while ignoring their perceived quality. The mean individuality scores of the examined methods for M2M, M2F, F2M and F2F conversions, and also their scores, averaged over all four conversions, are presented in Figures 5 and 6, respectively.

Without enhancement, the classical GMM conversions achieved similar scores, 5% higher than the proposed GB conversion. 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.

To conclude, applying GV enhancement significantly improves both quality and individuality of all the examined methods. Our proposed En-GB leads to comparable quality to the classical GMM-based conversion methods (without enhancement), and to comparable individuality to their enhanced versions.

5 Conclusion

We propose here a new method for spectral conversion, based on sequential Bayesian tracking, using a Grid-Based (GB) formulation. The target

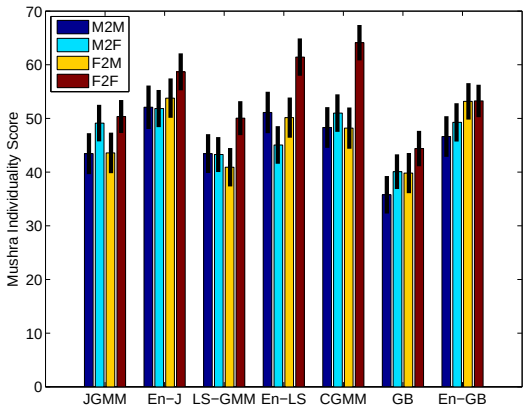

Fig. 5: Subjective individuality test, comparing: JGMM [16], Enhanced JGMM (En-J), LS-GMM [21], Enhanced LS-GMM (En-LS), CGMM [5], GB and Enhanced GB (En-GB).

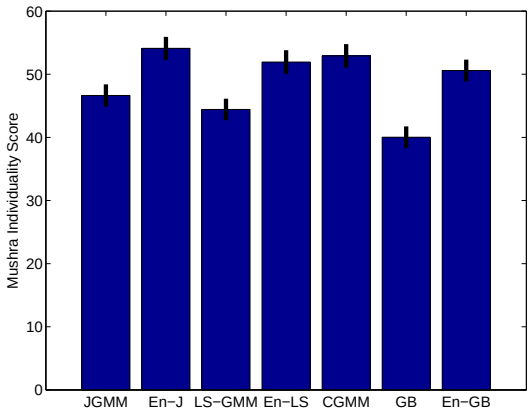

Fig. 6: Subjective individuality test averaged over all four gender conversions, comparing: JGMM [16], Enhanced JGMM (En-J), LS-GMM [21], Enhanced LS-GMM (En-LS), CGMM [5], GB and Enhanced GB (En-GB).

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 (using Matlab), 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.

During training, the evidence and likelihood probabilities needed for the GB formulation are approximated as discrete densities. During conversion, the converted spectrum is obtained as a weighted sum of the training target

vectors, used as grid-points. The weights are sequentially evaluated so that a smooth temporal evolution of the converted spectra is produced.

The GB conversion method enables the user to attain different NGV values by varying its two parameters (R_x, R_y). When tuned to minimal spectral distortion, GB achieves comparable objective performance to the classical GMM-based conversion methods. To further improve the quality of the synthesized speech, we increased the variability of the converted vectors by applying GV enhancement as a post-processing block.

We compared the proposed Enhanced GB (En-GB) scheme to CGMM and to classical GMM-based conversions, with and without GV enhancement, using listening tests. This comparison showed that En-GB achieves comparable quality to the classical GMM-based methods (without enhancement), and comparable individuality to their enhanced versions.

The proposed GB conversion, as most other methods, simply replaces the spectral envelopes extracted from the source signal with the converted outcome. As a result, the synthesized output has the same speaking rate as the source speaker. Further improvement can be obtained by modifying the duration of each converted utterance to match, on average, its corresponding value for the target speaker.

way for evaluating conversion systems. These objective measures may express significant trends and phenomena, but as shown here, they do not always agree with subjective evaluation results.

better correspondence to subjective results. In the mean time, subjective listening tests are imperative to properly evaluate and compare conversion methods.

The proposed GB conversion method, as presented here, is based on soft correspondence between the source and target vectors, obtained by using a parallel training set. Further research is needed to evaluate this correspondence for a non-parallel setup.

References

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

[2] M. R. Abe, S. Nakamura, K. Shikano, and H. Kuwabara. Voice conversion through vector quantization. In *Proc. ICASSP*, pages 655–658, 1988.

[3] B. Anderson and J. Moore. *Optimal Filtering*. Prentice-Hall, 1979.

[4] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Proc.*, 50(2):174–188, 2002.

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

[6] H. Benisty, D. Malah, and K. Crammer. Modular global variance enhancement for voice conversion systems. In *Proc. EUSIPCO*, pages 370–374, 2012.

[7] O. Cappe and E. Moulines. Regularization techniques for discrete cepstrum estimation. *IEEE Signal Processing Letters*, 3(4):100–102, 1996.

[8] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. R. Stat. Soc. B*, 39:1–38, 1977.

[9] T. En-Najjary, O. Rosec, and T Chonavel. A voice conversion method based on joint pitch and spectral envelope transformation. In *Proc. Interspeech ICSLP*, pages 1225–1225, 2004.

[10] D. Erro.

[11] D. Erro, A. Moreno, and A. Bonafonte. Inca algorythm for training voice conversion systems from nonparallel corpora. *IEEE Trans. on Audio, Speech and Lang. Proc.*, 18(5):944–953, 2010.

[12] D. Erro, A. Moreno, and A. Bonafonte. Voice conversion based on weighted frequency warping. *IEEE Trans. on Audio, Speech and Lang. Proc.*, 18(5):922–931, 2010.

[13] D. Erro, I. Sainz, and I. Hernaez. Improved HNM-based vocoder for statistical synthesizers. In *Proc. Interspeech*, pages 1809–1812, 2011.

[14] E. Godoy, O. Rosec, and T. Chonavel. Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora. *IEEE Trans. on Audio, Speech and Lang. Proc.*, 20(4):1313–1323, 2012.

[15] E. Helander, J. Schwarz, Silen H. Nurminen, J, and M. Gabbouj. On the impact of alignment on voice conversion performance. In *Proc. Interspeech*, pages 1453–1456, 2008.

[16] A. Kain and M. Macon. Spectral voice conversion for text-to-speech synthesis. In *Proc. ICASSP*, pages 285–288, 1998.

[17] A. Kain and M. W. Macon. Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction. In *Proc. ICASSP*, pages 813–816, 2001.

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

[19] H. Kuwabara and Y Sagisaka. Acoustic characteristics of speaker individuality: Control and conversion. *IEEE Transactions on Signal Proc.*, 16(2):165–173, 1995.

[20] L. Mesbashi, V. Barreaud, and O. Boeffard. Comparing GMM-based speech transformation systems. In *Proc. Interspeech*, pages 1456–1989, 2007.

[21] O. Stylianou, Y. Cappe and E. Moulines. Continuous probabilistic transform for voice conversion. *IEEE Trans. on Speech and Audio Proc.*, 6(2):131–142, 1998.

[22] T. Toda, A. W. Black, and K. Tokuda. Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter. In *Proc. ICASSP*, pages 9–12, 2005.

[23] T. Toda, A. W. Black, and K. Tokuda. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Trans. on Audio, Speech and Lang. Proc.*, 15(8):2222–2235, 2007.

[24] T. Toda, Y. Ohtani, and K. Shikano. Eigenvoice conversion based on Gaussian mixture model. In *Proc. ICSLP*, pages 2446–2449, 2006.

[25] T. Toda, H. Saruwatari, and K. Shikano. Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum. In *Proc. ICASSP*, pages 841–844, 2001.

[26] H. Ye and S. Young. Quality-enhanced voice morphing using maximum likelihood transformations. *IEEE Trans. on Audio, Apeech and Lang. Proc.*, 14(4):1301–1312, 2006.