SIPL SIGNAL & IMAGE PROCESSING LAB

Sequential Voice Conversion Using Grid-Based Approximation

H. Benisty, D. Malah and K. Crammer

Electrical Engineering Department

Technion – Israel Institute of Technology

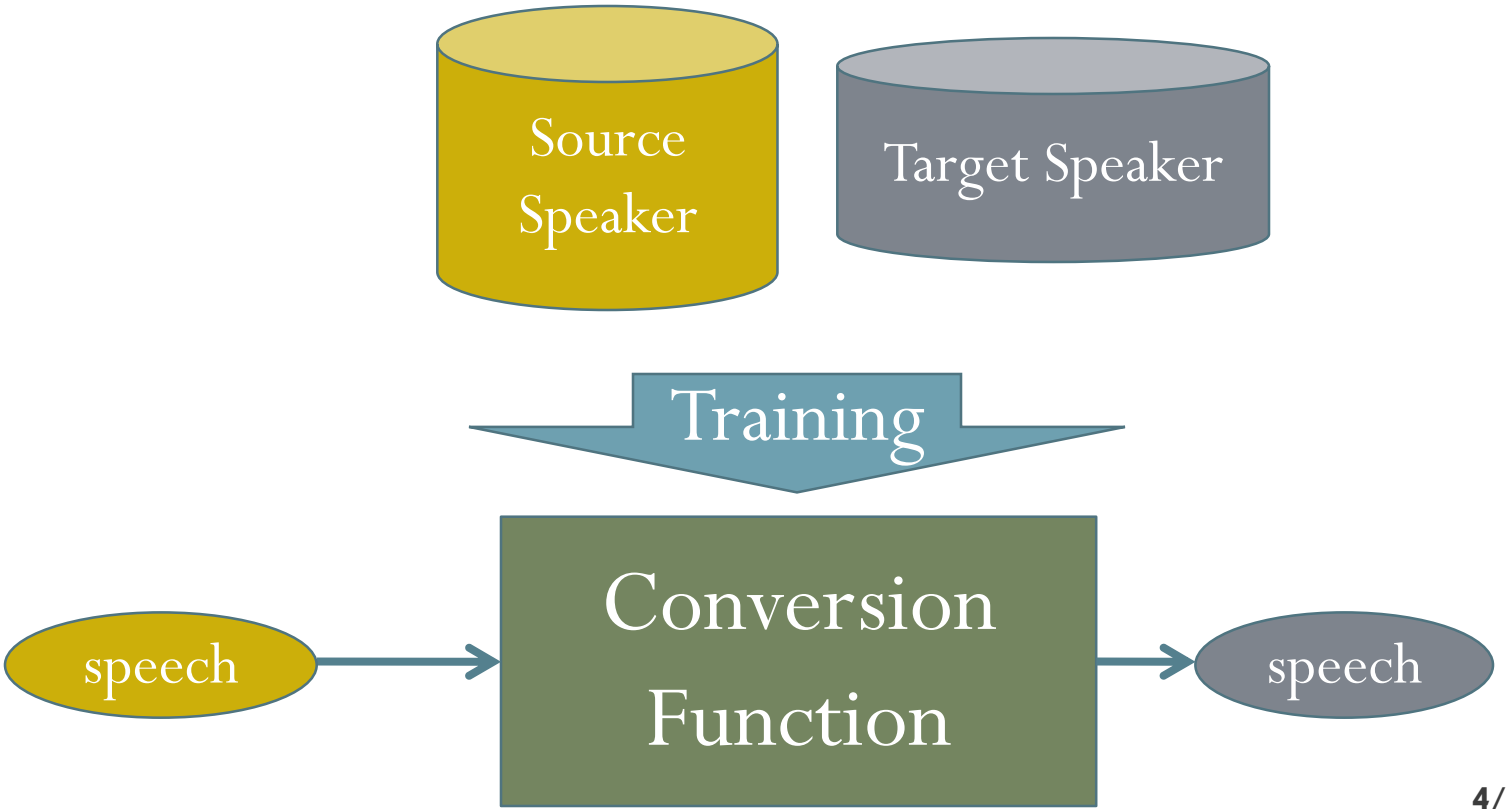
IEEEI 2014 – Eilat, Israel

Agenda

- Introduction
  - Motivation
  - General VC Scheme
- A Common Method - GMM-based conversion
- Grid-Based (GB) Conversion
  - Formulation
- Experimental Results
- Conclusion

General Conversion Setup

- **The goal**: modify a source speaker's speech to sound as if spoken by a target speaker



Applications

- Personalize the output of Text-to-Speech (TTS) systems
- Dialog systems
- Vocal pathology
- Entertainment
- Toys & games

Speech Characteristics

- The identity of a speaker is associated with:
  - Prosody attributes - pitch, duration and energy
  - Spectral envelope
- Pitch - usually modified using a simple statistical mean and variance scaling
- Most VC methods deal with **spectral envelope conversion**

Training Data

- Textual Content
  - Parallel data set – the two speakers say the same text
  - Non-parallel data set – source-target correspondence is learned during training
  - Cross-lingual data set

GMM-Based Conversion

[Stylianou et. al., 1998; Kain & Macon, 1998]

- Given a **parallel** and **aligned** source and target training vectors $\left\{\mathbf{x}^k, \mathbf{y}^k\right\}_1^N \in \Re^P$ (represented by Mel Frequency Cepstrum Coefficients - MFCCs)
- A GMM is trained using the source vectors:

$$p(\mathbf{x}) = \sum_{m=1}^{M} \alpha_m N\left(\mathbf{x}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\right)$$

- The conversion function is formulates, using the GMM parameters, as a weighted sum of linear Bayesian estimators of the target spectra:

$$\mathcal{F}(\mathbf{x}) = \sum_{m=1}^{M} \alpha_m \left(A_m \mathbf{x} + b_m\right)$$

GMM-Based Conversion - Cont.

- Exhaustive training - Expectation Maximization
  - Often leading to ill-conditioning if the dataset used is too small
- Linear conversion → over-smoothed spectral envelops → muffled synthesizes speech
  - Proposed remedy – global variance enhancement

    [Toda et al, 2007; Benisty and Malah, 2011; Benisty et al., 2012]

Grid-Based (GB) Conversion

- Based on sequential **Bayesian tracking**
- 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**

Grid-based conversion

Bayesian Tracking

- The target spectrum $\mathbf{y}_t$, is assumed to follow a first order Markov dynamics: $\mathbf{y}_t = f_t(\mathbf{y}_{t-1}, \mathbf{u}_t)$

  $\mathbf{u}_t$ - an i.i.d. noise sequence.

- The source spectrum $\mathbf{x}_t$, depends on the target spectrum through: $\mathbf{x}_t = h_t(\mathbf{y}_t, \mathbf{v}_t)$

  $\mathbf{v}_t$ - an i.i.d. measurement noise

- The Bayesian optimal estimation for the target spectrum is:

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

- In practice an analytical derivation is impossible since we do not have a closed form for $p\left(\mathbf{y}_t \middle| \mathbf{x}_{1:T}\right)$, so we use **Grid-Based approximation**

Grid-based conversion

Discrete Approximation

- Given: parallel unaligned training sets – $\left\{\mathbf{x}^k\right\}_1^{N_x} \in \Re^P, \left\{\mathbf{y}^k\right\}_1^{N_y} \in \Re^P$ and a sequence of test source vectors - $\mathbf{x}_{1:T}$
- We evaluate the posterior probability as a discrete sum:

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

where the posterior weights are: $w_{t|t}^k = p\left(\mathbf{y}_t = \mathbf{y}^k \left| \mathbf{x}_{1:T} \right.\right)$

- The optimal Bayesian estimation for the target spectrum is evaluated as a discrete sum of the target training vectors:

$$\Rightarrow \boxed{\hat{\mathbf{y}}_t = \sum_{k=1}^{N_y} w_{t|t}^k \mathbf{y}^k}$$

Grid-based conversion

Sequential Estimation

- The posterior weights are sequentially evaluated using two stages:

1. Prediction: $$w_{t|t-1}^{k} \approx \sum_{l=1}^{N_y} w_{t-1|t-1}^{l} p\left(\mathbf{y}_t = \mathbf{y}^k \middle| \mathbf{y}_{t-1} = \mathbf{y}^l\right)$$

2. Update: $$w_{t|t}^{k} \approx \frac{w_{t|t-1}^{k} p\left(\mathbf{x}_t \middle| \mathbf{y}_t = \mathbf{y}^k\right)}{\sum_{l=1}^{N_y} w_{t|t-1}^{l} p\left(\mathbf{x}_t \middle| \mathbf{y}_t = \mathbf{y}^l\right)}$$

- $p\left(\mathbf{y}_t = \mathbf{y}^k \middle| \mathbf{y}_{t-1} = \mathbf{y}^l\right)$ - evidence probability
- $p\left(\mathbf{x}_t \middle| \mathbf{y}_t = \mathbf{y}^k\right)$ - likelihood probability

Grid-based conversion

Evidence Modeling

- The evidence probability $p\left(\mathbf{y}_t = \mathbf{y}^k \middle| \mathbf{y}_{t-1} = \mathbf{y}^l\right)$ expresses the transition probability from state $\mathbf{y}^l$ to $\mathbf{y}^k$
- In natural speech, spectral feature vectors related to consecutive time frames are typically similar, but not identical.
- We model the transition probability as:

$$p\left(\mathbf{y}_t = \mathbf{y}^k \middle| \mathbf{y}_{t-1} = \mathbf{y}^l\right) \propto \exp\left\{-\frac{1}{2}\max\left(\frac{\mathrm{MCD}\left(\mathbf{y}^k, \mathbf{y}^l\right)}{R_y}, 1\right)^2\right\}$$

  - R_y – a parameter
- MCD – Mel Cepstral Distortion:

$$\mathrm{MCD}\left(\mathbf{y}^k, \mathbf{y}^l\right) \triangleq \frac{10\sqrt{2}}{\ln 10}\left\|\mathbf{y}^k - \mathbf{y}^l\right\|_2$$

Grid-based conversion

Likelihood Modeling

- The likelihood probability $p\left(\mathbf{x}_t \left| \mathbf{y}_t = \mathbf{y}^k \right.\right)$ expresses the probability of a source vector given target training vectors
- We model the likelihood probability as:

$$p\left(\mathbf{x}_t \left| \mathbf{y}_t = \mathbf{y}^k \right.\right) \propto \sum_{m=1}^{N_x} p\left(\mathbf{x}^m \left| \mathbf{y}_t = \mathbf{y}^k \right.\right) \exp\left\{ -\frac{\mathrm{MCD}^2\left(\mathbf{x}_t, \mathbf{x}^m\right)}{2R_x^2} \right\}$$

  - R_x – a parameter
  - $p\left(\mathbf{x}^m \left| \mathbf{y}_t = \mathbf{y}^k \right.\right)$ - the discrete likelihood

Grid-based conversion

Discrete Likelihood Modeling

- The discrete likelihood $p\left(\mathbf{x}^m \middle| \mathbf{y}_t = \mathbf{y}^k\right)$ expresses the correspondence between the source and target training vectors
- We use a parallel (but need not be aligned) and phonetically labeled data set to model the discrete likelihood as:

$$p\left(\mathbf{x}^m \middle| \mathbf{y}_t = \mathbf{y}^k\right) \propto \begin{cases} 1 & \mathbf{x}^m \text{and } \mathbf{y}^k \text{belong to the same mid-utterance} \\ 0 & \text{otherwise} \end{cases}$$

Grid-Based (GB) Conversion Algorithm Summary

- **Input:** a sequence of feature vectors
- **Initialization:** set the initial weights, $\left\{ w_{t|t-1}^{0} \right\}_{k=1}^{N_y}$
- **Main Iteration: for t = 1, ...T , perform the following steps:**

1. Evaluate the prior weights: $\left\{ w_{t|t-1}^{k} \right\}_{k=1}^{N_y}$
2. Evaluate the posterior weights: $\left\{ w_{t|t}^{k} \right\}_{k=1}^{N_y}$
3. Obtain the converted spectra:

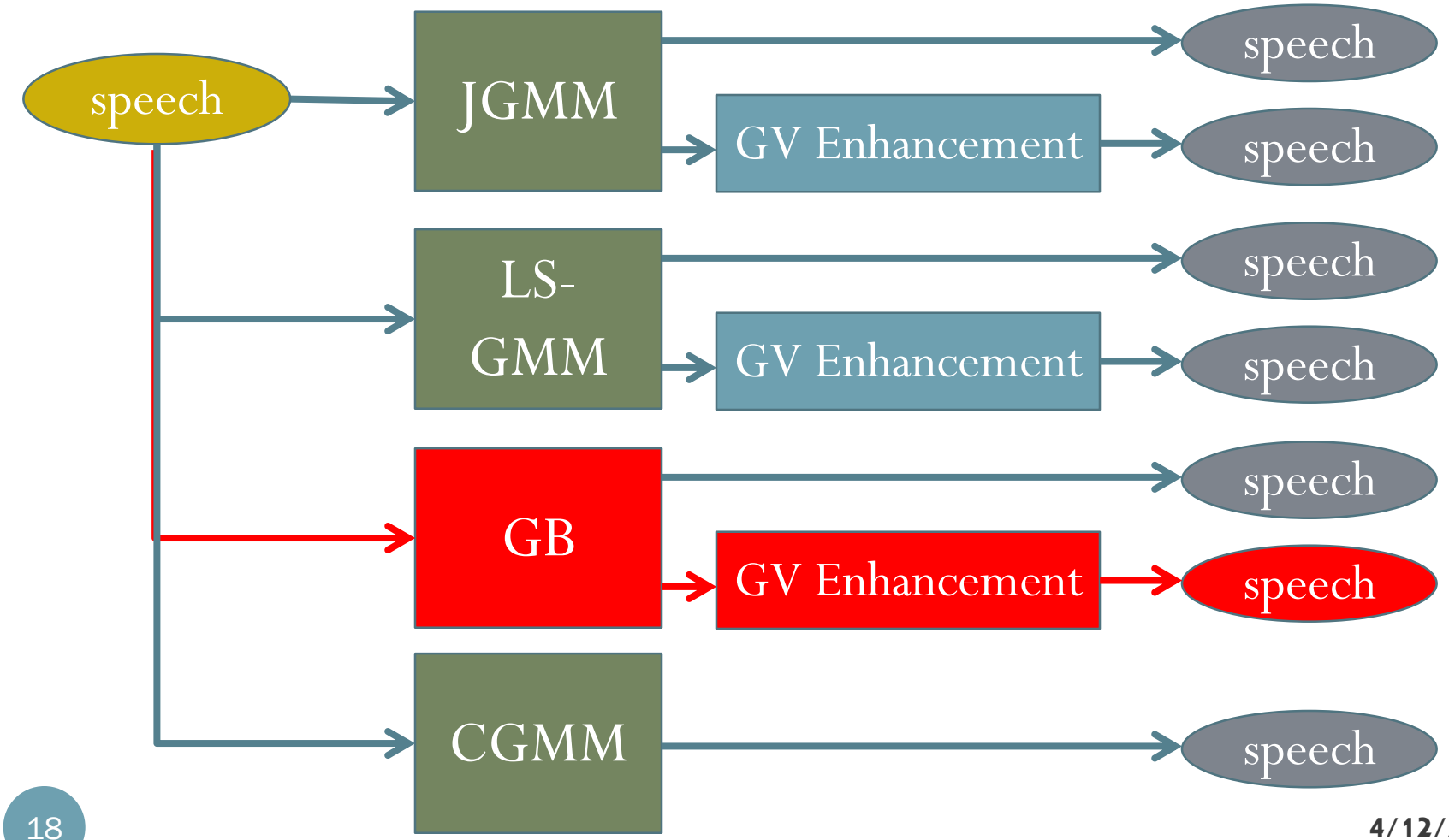
$$\hat{\mathbf{y}}_t = \sum_{k=1}^{N_y} w_{t|t}^{k} \mathbf{y}^k$$

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

Examined Methods

- Conversion Methods
  - Joint GMM (JGMM) [Kain & Macon, 1998]
  - Least squares GMM (LS-GMM) [Stylianou et. al., 1998]
  - Proposed GB
- GV-Enhancement Methods
  - Integrated during training –
    Constrained GMM (CGMM) [Benisty and Malah, 2011]
  - A Post processing block [Benisty et. al., 2012]

Examined Setups

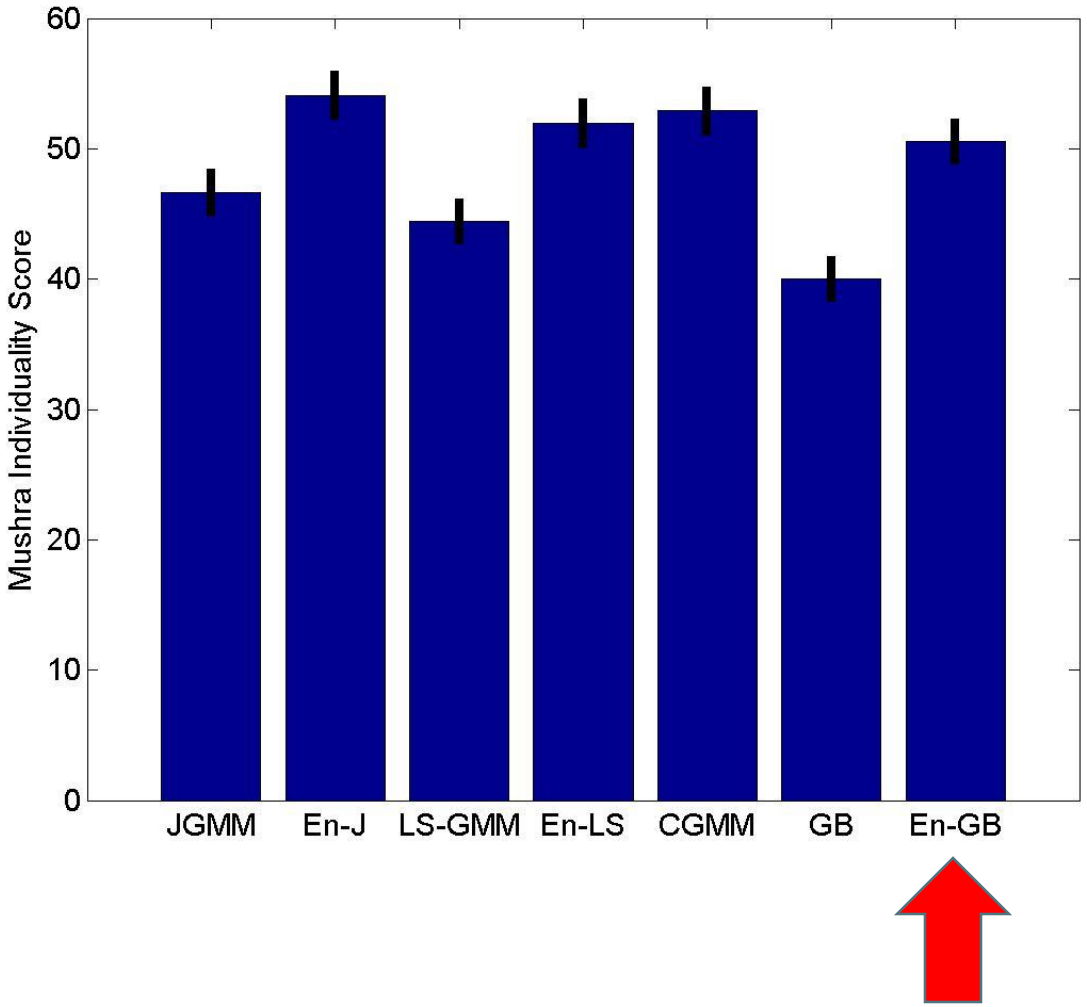


Training Period

Method	Training time	Conversion time per frame
JGMM	7 h	11 msec
LS-GMM	8.5 h	11 msec
CGMM	11 h	11 msec
GB	**10 sec**	10 msec
GB enhancement	None in training	23 msec

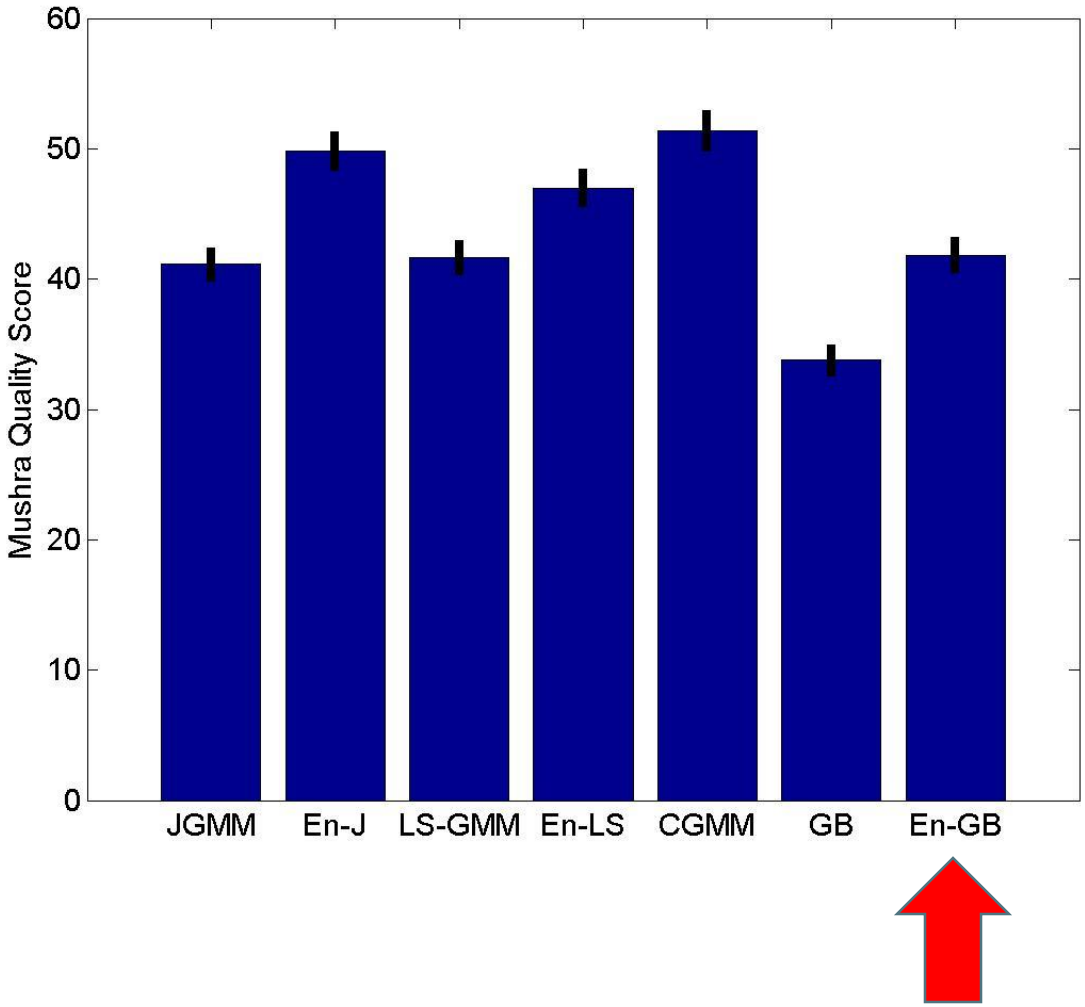
Subjective Evaluations

Individuality Tests



Subjective Evaluations

Quality Tests



Conclusion

- Easily trained using either small or large scale datasets
- Does require a parallel dataset but **without** time alignment
- Subjectively –
  - Comparable quality to the classical GMM-based methods (without enhancement)
  - Comparable individuality enhanced GMM-based methods
- Sound samples are available at: http://sipl.technion.ac.il/Info/hadas/sound-samples.htm

Thank You