MODULAR GLOBAL VARIANCE ENHANCEMENT FOR VOICE CONVERSION SYSTEMS

Hadas Benisty, David Malah, and Koby Crammer

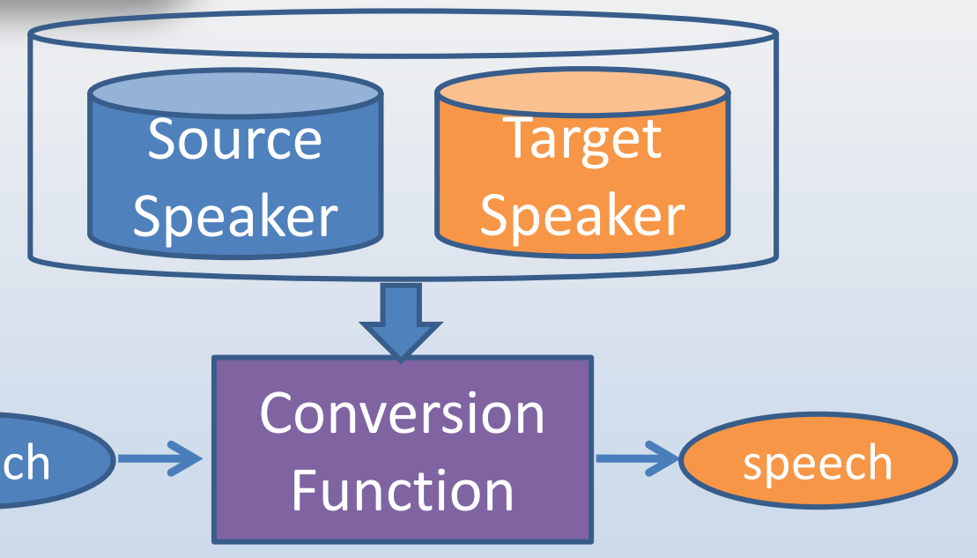
EE Department, Technion – Israel Institute of Technology, Haifa, Israel

• Goal

- ❑ Many voice conversion methods produce muffled synthesized outputs due to over-smoothing of the converted spectra
- ❑ GV enhancement – used for muffling reduction and commonly applied as an integrated part of the conversion system
- ❑ **We propose a new modular method for GV enhancement, applied as a post-processing block**

• Voice Conversion

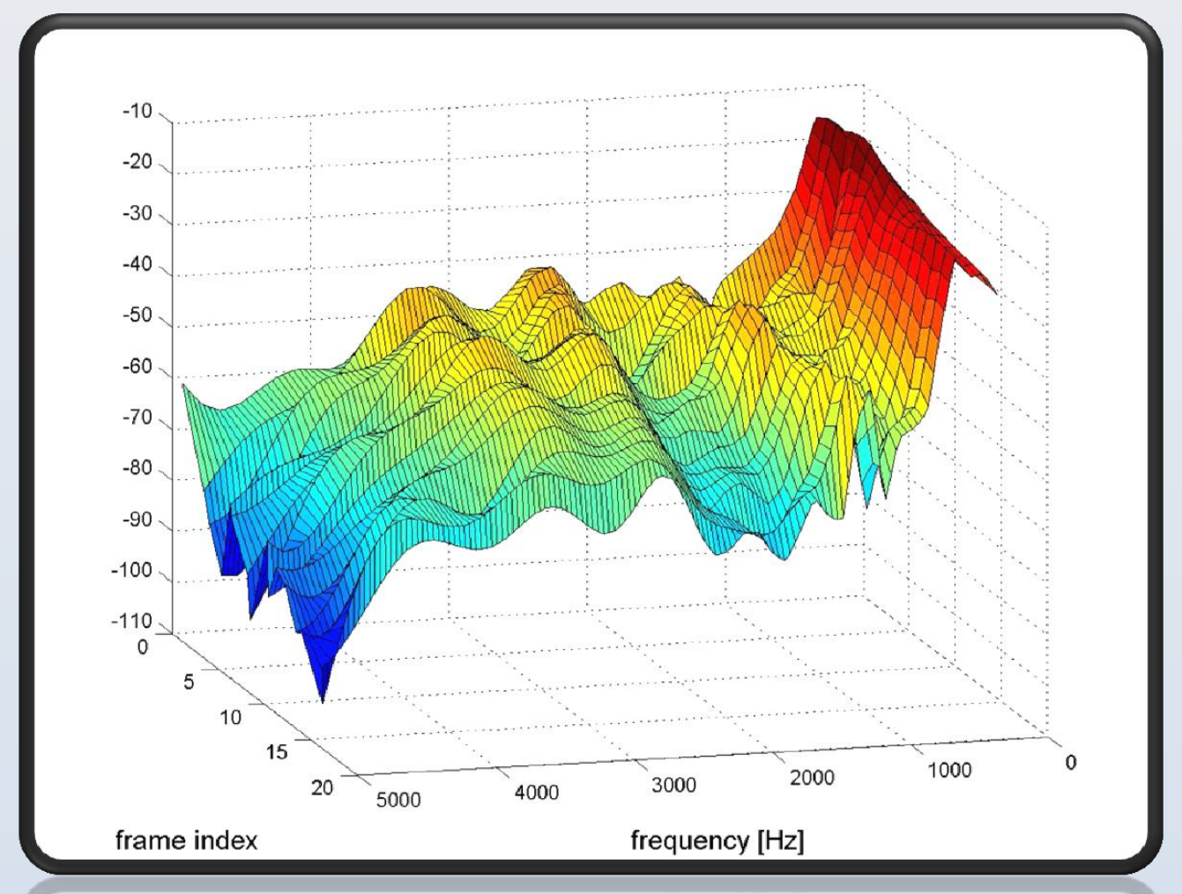
- ❑ Transform a sentence said by a source speaker, to sound as if a target speaker had said it, based on pre-recorded training set



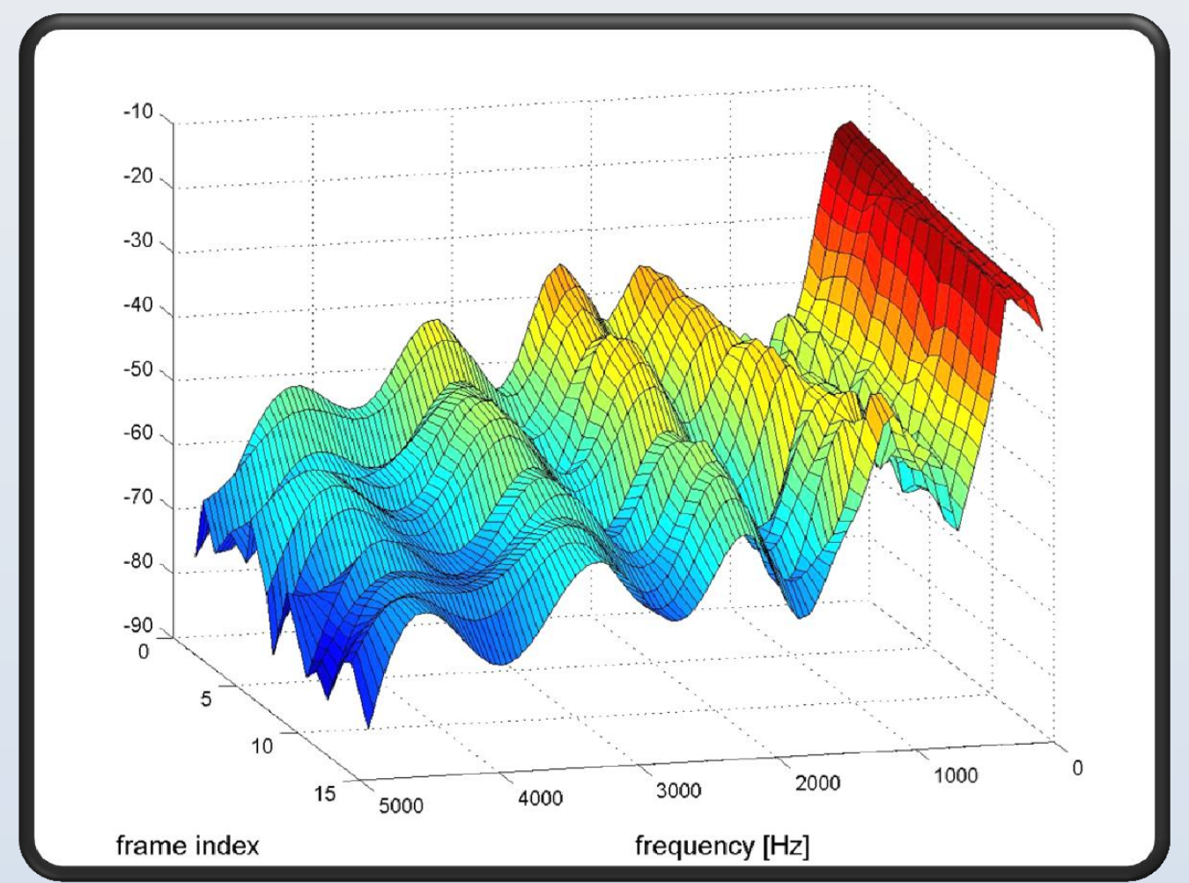
• Voice Conversion Using GMM

- ❑ Linear Conversion based on a Gaussian Mixture Model (GMM) [Stylianou, 1998], [Kain & Macon, 1998]
- ❑ A common approach for spectral conversion
- ❑ Minimizes the mean Log Spectral distortion (LSD) between converted feature vectors and target vectors
- ❑ Characterized by smoothed spectral envelopes causing a **muffling effect:**

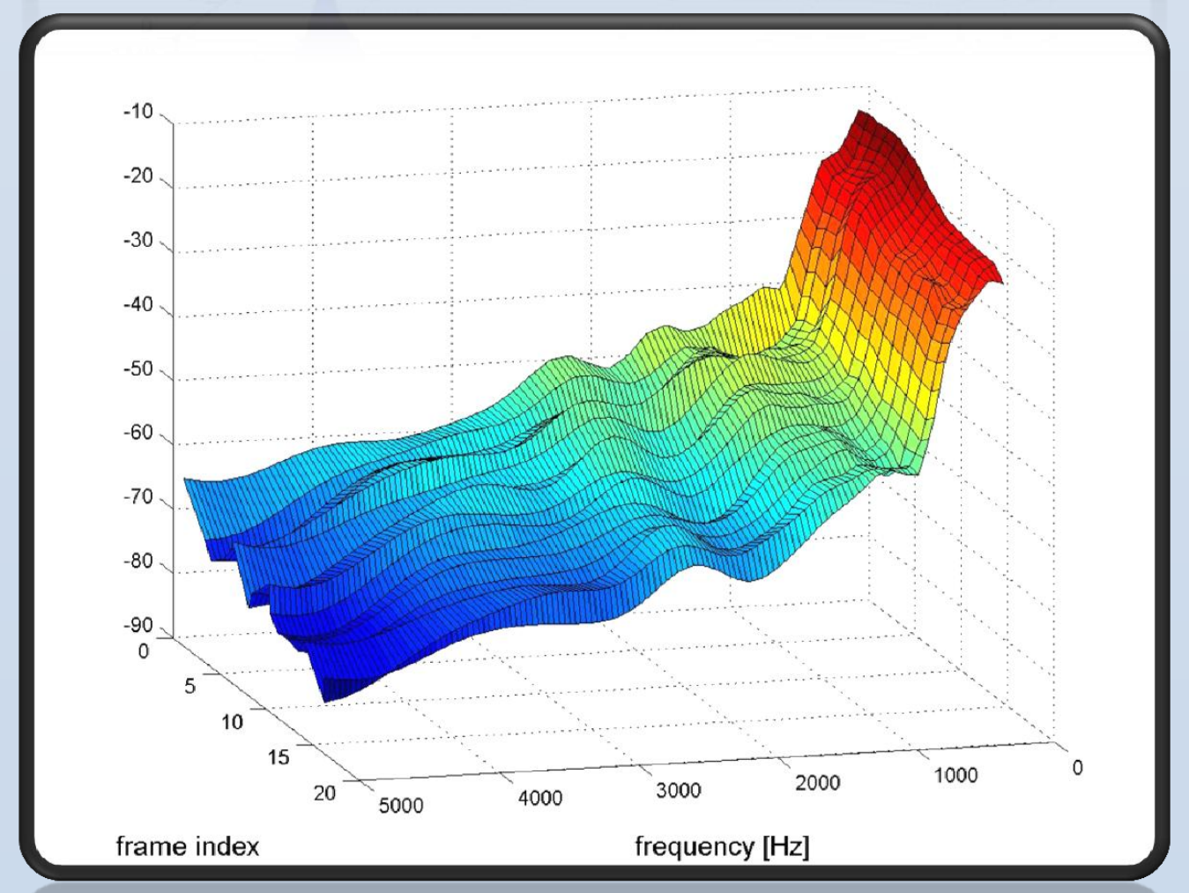
Spectral-Envelope Evolution in Time



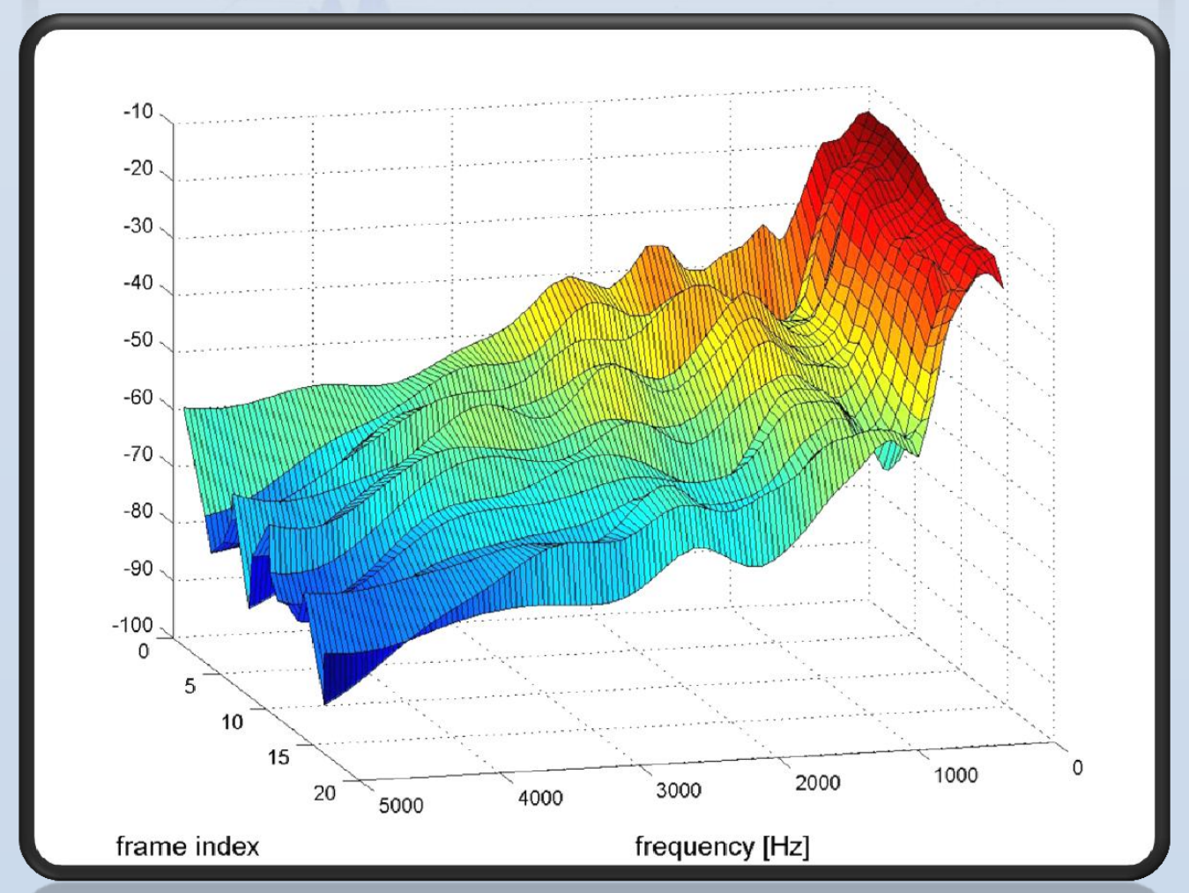
Source Speaker



Target Speaker



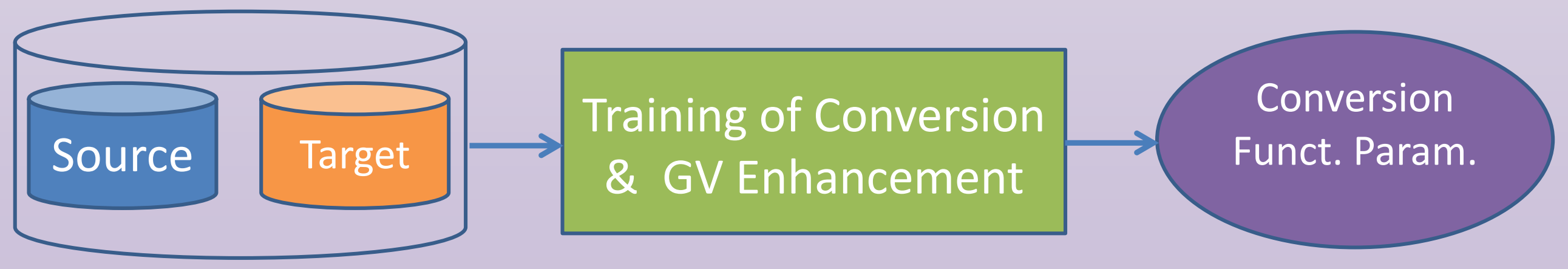
Converted signal [Stylianou, 1998]



LS-GMM followed by GV enhancement **(our work)**

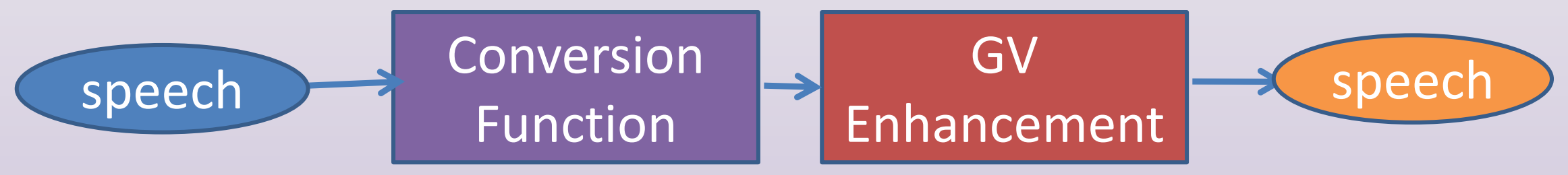
• GV Enhancement

- ❑ GV enhancement methods have been proposed to overcome the muffling effect:
  - ○ ML estimation [Toda et. al., 2007]
  - ○ Constrained GMM (CGMM) [Benisty and Malah, 2011]
- ❑ These enhancement methods are integrated into the training process of the conversion



• Proposed Modular GV Enhancement

- ❑ **GV Enhancement Using an LSD Constraint**
  - ○ Designed **independently** of any specific conversion scheme and applied as a **post-processing** block



  - ○ The extent of GV enhancement is **controlled** by the allowed **spectral distance** the enhanced and the originally converted output, as specified by the user

• GV Enhancement Using an LSD Constraint

- ❑ **Input**
  - ○ A sequence of converted feature vectors $\tilde{\mathbf{Y}}_{1:T} \triangleq (\tilde{\mathbf{y}}_1, \quad \tilde{\mathbf{y}}_2, \quad \ldots, \quad \tilde{\mathbf{y}}_T)^T$
- ❑ **Output**
  - ○ A sequence of **enhanced** feature vectors $\tilde{\mathbf{Z}}_{1:T} \triangleq (\tilde{\mathbf{z}}_1, \quad \tilde{\mathbf{z}}_2, \quad \ldots, \quad \tilde{\mathbf{z}}_T)^T$
- ❑ The enhanced sequence is the solution of:

$$\tilde{\mathbf{Z}}_{1:T} = \underset{\mathbf{Z}_{1:T}}{\arg\max}\, \mathrm{NGV}\{\mathbf{Z}_{1:T}\} \quad \text{s.t} \quad \overline{\mathrm{LSD}}\left(\mathbf{Z}_{1:T}, \tilde{\mathbf{Y}}_{1:T}\right) \le \theta_{LSD}$$

  - ○ $\mathrm{NGV}\{\mathbf{Z}_{1:T}\}$ - the normalized GV of the sequence $\mathbf{Z}_{1:T}$, evaluated by:

$$\mathrm{NGV}\{\mathbf{Z}_{1:T}\} \triangleq \frac{1}{P}\sum_{p=1}^{P} \frac{\mathrm{Var}\left\{\tilde{\mathbf{Z}}_{1:T}(p)\right\}}{\mathrm{Var}\left\{\mathbf{Z}_{1:T}(p)\right\}}$$

  - ○ $\overline{\mathrm{LSD}}\left(\mathbf{Z}_{1:T}, \tilde{\mathbf{Y}}_{1:T}\right)$ - mean Log spectral Distortion between the converted and enhanced sequences
  - ○ θ_{LSD} - pre-set threshold value for the mean LSD in dB
- ❑ The solution is obtained with explicit terms for mean LSD and NGV

$$\overline{\mathrm{LSD}}\left(\tilde{\mathbf{Z}}_{1:T}, \tilde{\mathbf{Y}}_{1:T}\right) \approx \frac{\kappa}{T}\left\|\tilde{\mathbf{Z}}_{1:T} - \tilde{\mathbf{Y}}_{1:T}\right\|_{2,1} \qquad \kappa \triangleq 10\sqrt{2}/\ln 10$$

$$\mathrm{NGV}\left\{\tilde{\mathbf{Y}}_{1:T}\right\} = \frac{1}{P}\left\|\Delta_T \cdot \tilde{\mathbf{Y}}_{1:T} \cdot \mathbf{C}^{-\frac{1}{2}}\right\|_2^2 \qquad \Delta_T \triangleq \frac{1}{\sqrt{T}}\left(\mathbf{I}_{T\times T} - \frac{1}{T}\mathrm{ones}(T,T)\right)$$

$$\mathbf{C} \triangleq diag\left(\mathrm{Var}\{\mathbf{Y}(1)\}, \quad \ldots, \quad \mathrm{Var}\{\mathbf{Y}(P)\}\right)$$

$\mathrm{Var}\{\mathbf{Y}(p)\}$ - GV of spectral features related to the target speaker

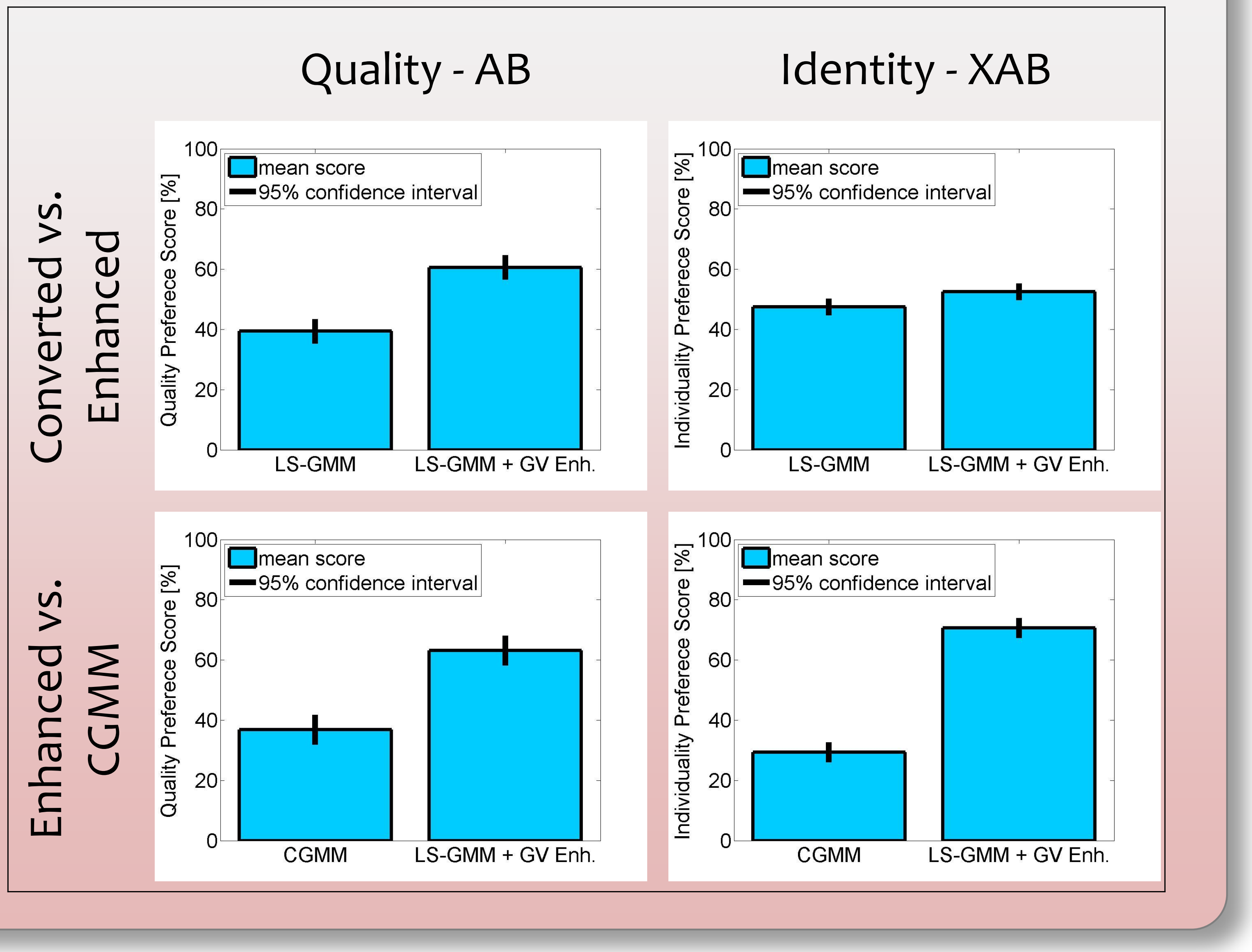
• Experimental Results

- ❑ **Evaluated Methods**
  - ○ GMM-based Conversion (LS-GMM) [Stylianou, 1998]
  - ○ LS-GMM followed by **our** GV enhancement
  - ○ CGMM [Benisty and Malah, 2011]
- ❑ **Objectively**
  - ○ For a given mean LSD, CGMM leads to higher GV than our method
- ❑ **Subjectively**
  - ○ Our method was selected by the majority of listeners as **better** than CGMM, both in terms of **quality** and **similarity** to the target

Objective Evaluations

Conversion Method	Mean LSD [dB]	Mean Norm. GV
LS-GMM	6.2	0.1
Enhanced $\theta_{LSD} = 1dB$	6.4	0.2
Enhanced $\theta_{LSD} = 2dB$	6.7	0.3
Enhanced $\theta_{LSD} = 4dB$	7.3	0.4
CGMM	7.3	0.9

Subjective Evaluations


Quality - AB
Identity - XAB
Converted vs. Enhanced
Enhanced vs. CGMM
Quality Preferece Score [%]
Individuality Preferece Score [%]
mean score
95% confidence interval
LS-GMM
LS-GMM + GV Enh.
CGMM