# Voice Conversion using GMM with Enhanced Global Variance

Hadas Benisty and David Malah

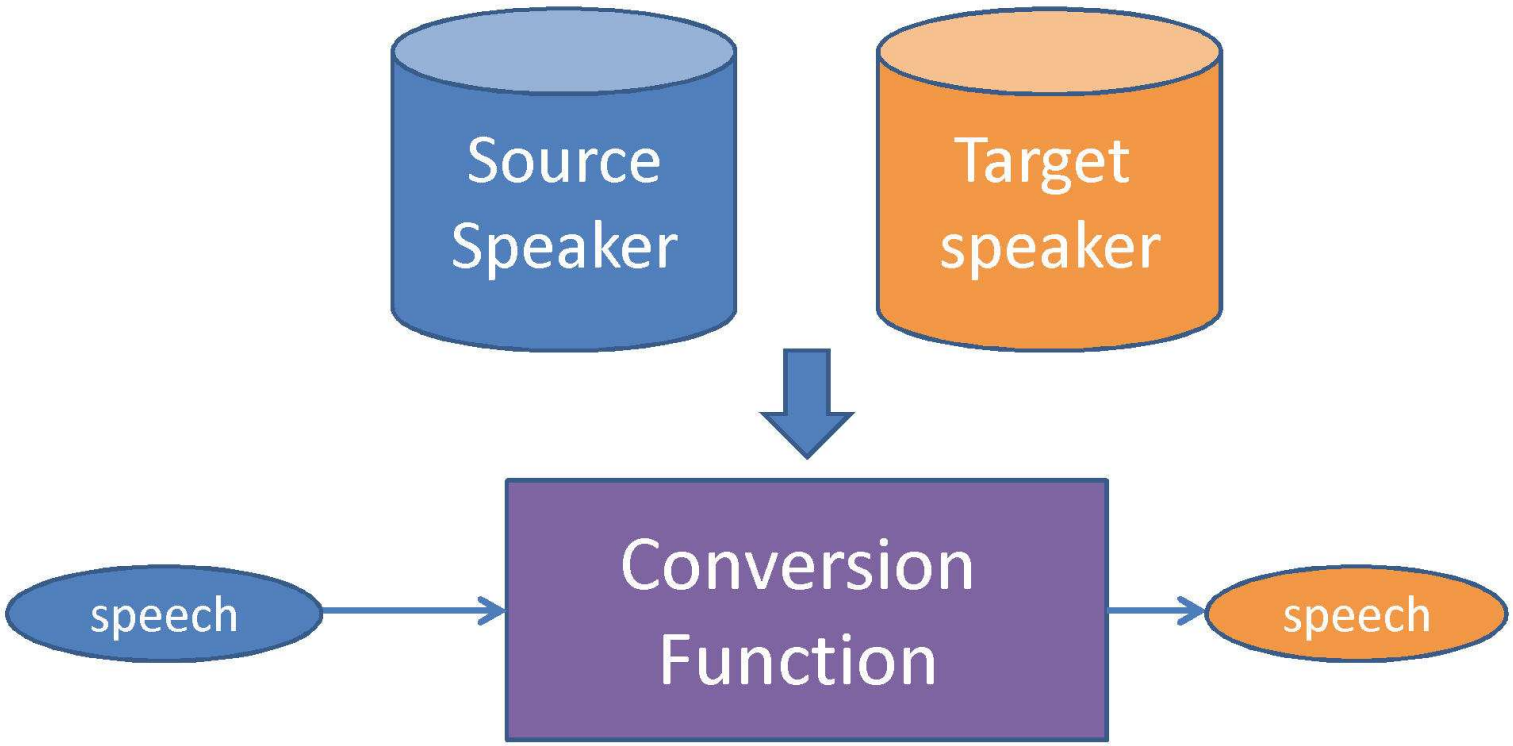
Electrical Engineering Department
Technion – Israel Institute of Technology

Interspeech 2011
Florence, Italy

# Outline

1. **Introduction**
   - Motivation
   - General VC Scheme
   - Common Methods
2. **GMM-based Conversion**
   - General Concept
3. **Constrained GMM**
   - Formulation
4. **Simulation Results**
   - Setup
   - Results
5. **Conclusion**

# Voice Conversion

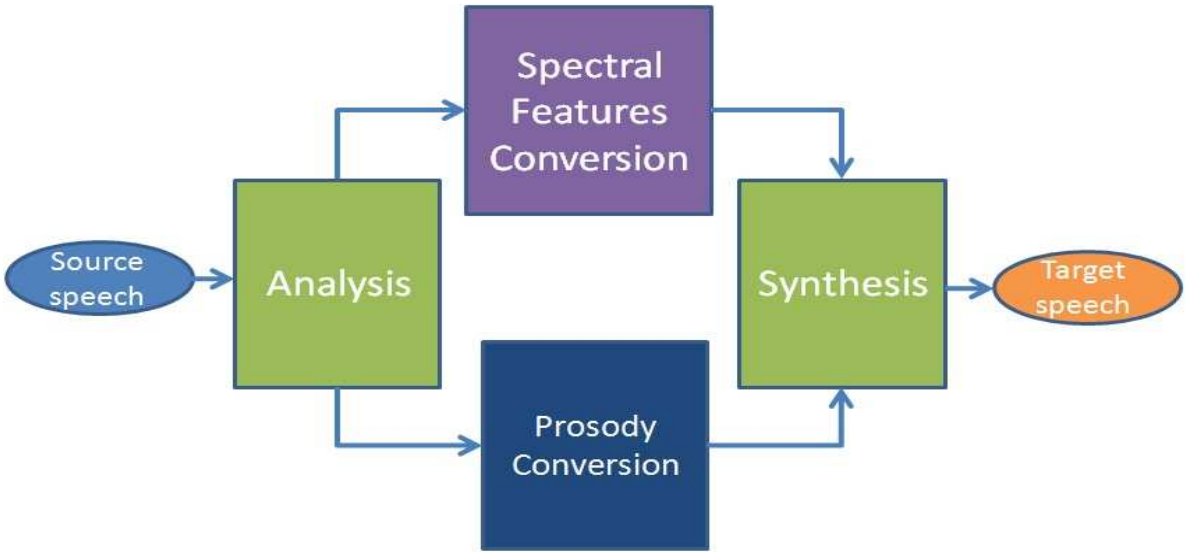



- **The goal** – transform a sentence said by a source speaker, to sound as if a target speaker had said it, based on pre-recorded training set

# Applications

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

# Common Voice Conversion Scheme



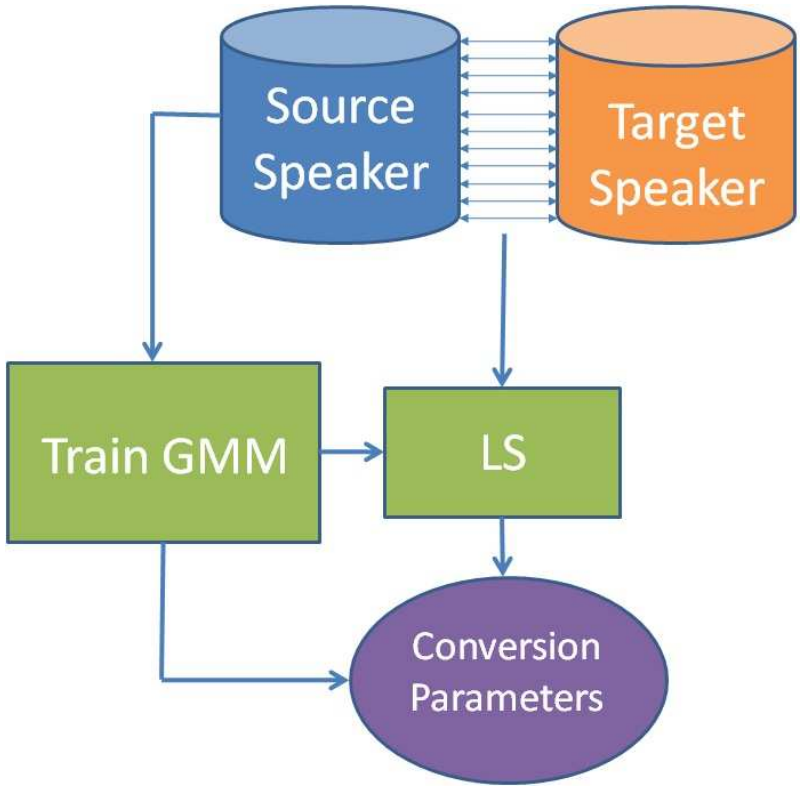

- The synthesis is performed after the conversion of:
  - Spectral features
    - Line Spectral Frequencies (LSFs) (Kondoz, 1994)
    - Mel Frequency Cepstrum Coefficients (MFCCs) (Cappe et. al. 1995)
  - Prosody features - pitch, duration and energy

# Spectral Envelope Conversion Approaches

- **Gaussian Mixture Model (GMM) and Linear Conversion**
  - [Stylianou, 1998]; [Kain & Macon, 1998]
- **Modifications of the GMM-Based Conversion**
  - [Toda, 2001]; [Kain and Macon, 2001]; [En-Najjary et. al., 2004]; [Helaner et. al., 2010]; [Erro et. al., 2010]
- **Codebook Selection**
  - [Abe, 1998]; [Arslan, 1999]; [Suderman, 2005]
- **Hidden Markov Models (HMM)**
  - [Ye and Young, 2006]; [Zhang et. al., 2008]

# GMM-based Conversion Trained by LS
**[Stylianou, 1999]**

- Time alignement using DTW
- GMM training
- A linear conversion function - defined using the GMM parameters
- Conversion parameters - evaluated using a Matrix formulation and LS



# Constrained Conversion - Motivation

- GMM-based conversion
  - Minimizes the mean LSD between the converted feature vectors and the target vectors
  - Characterized by smoothed spectral envelopes causing a muffling effect:

    source  converted output  target 
  - Toda et. al., 2007 suggested increasing the global variance (GV) of the spectral features using ML estimation

**Proposed Solution:**

Minimize the mean LSD, while the GV of the converted output is **constrained** to match the GV of the target speaker

# Constrained Formulation

## Constrained Training

$$\min_{\mathbf{q}^p} \|\mathbf{A}^p\mathbf{q}^p - \mathbf{b}^p\|^2$$
$$s.t. \quad \|\mathbf{B}^p\mathbf{q}^p\|^2 = c^2(p) \qquad p = 1, ..., P$$

| Notation | Definition | Dimmension |
|---|---|---|
| $\mathbf{B}^p$ | $\mathbf{\Delta A}^p$ | $Q \times 2M$ |
| $c^2(p)$ | $\|\mathbf{\Delta y}^p\|^2$ | $1 \times 1$ |

$$\mathbf{\Delta} \triangleq \frac{1}{\sqrt{Q}} \left( \mathbf{I}_{Q\times Q} - \frac{1}{Q} \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \cdots & \vdots \\ 1 & \cdots & 1 \end{pmatrix} \right)$$

- The $P$ constrained minimization problems can be solved by using the Lagrange Multiplier method and joint diagonalization of the pairs $\{\mathbf{A}^p, \mathbf{B}^p\}_{p=1}^P$

# Data Structure

- Two U.S. English male speakers from the CMU ARCTIC database (by Festvox)
- 50 parallel training sentences
- 50 parallel testing sentences
- Analysis and synthesis - using the Harmonic Plus Noise Model (HNM) [Stylianou, 2001]
- Spectral features - 24 MFCC's

# Conversion Scheme

- Spectral Conversion -
  - The GV of the first 12 MFCC's was constrained
  - The last 12 coefficients were not constrained
- Pitch
  - Linear conversion of the global statistics:

$$\hat{f}_0^{(y)}(k) = \mu^{(y)} + \frac{\sigma^{(y)}}{\sigma^{(x)}}\left(f_0^{(x)}(k) - \mu^{(x)}\right),$$

  - $f_0^{(x)}(k)$, $\hat{f}_0^{(y)}(k)$ - pitch values of the source and converted signals at the $k$-th frame
  - $\mu^{(x)}$, $\mu^{(y)}$ - mean pitch values
  - $\sigma^{(x)}$, $\sigma^{(y)}$ - standard deviations of the source and target pitch values

# Demonstration

| | Source | Target | Classical GMM | Constrained GMM |
|---|---|---|---|---|
| 1 | | | | |
| 2 | | | | |
| 3 | | | | |

*More samples are available at:
http://sipl.technion.ac.il/Info/hadas/sound-samples.htm

# Objective Measures

- **Similarity** to the target - mean Log Spectral Distortion (LSD):

$$LSD\left(\mathcal{F}\{\mathbf{x}\},\mathbf{y}\right) \approx \frac{10}{ln10}\sqrt{2\sum_{p=1}^{P}\left\|\mathcal{F}\{x_p\}-y_p\right\|^2}$$

- Mean normalized GV -

$$\text{Mean Norm. GV} = \frac{1}{P_0}\sum_{p=1}^{P_0}\frac{\text{Converted GV}\,(p)}{\text{Target GV}\,(p)}$$

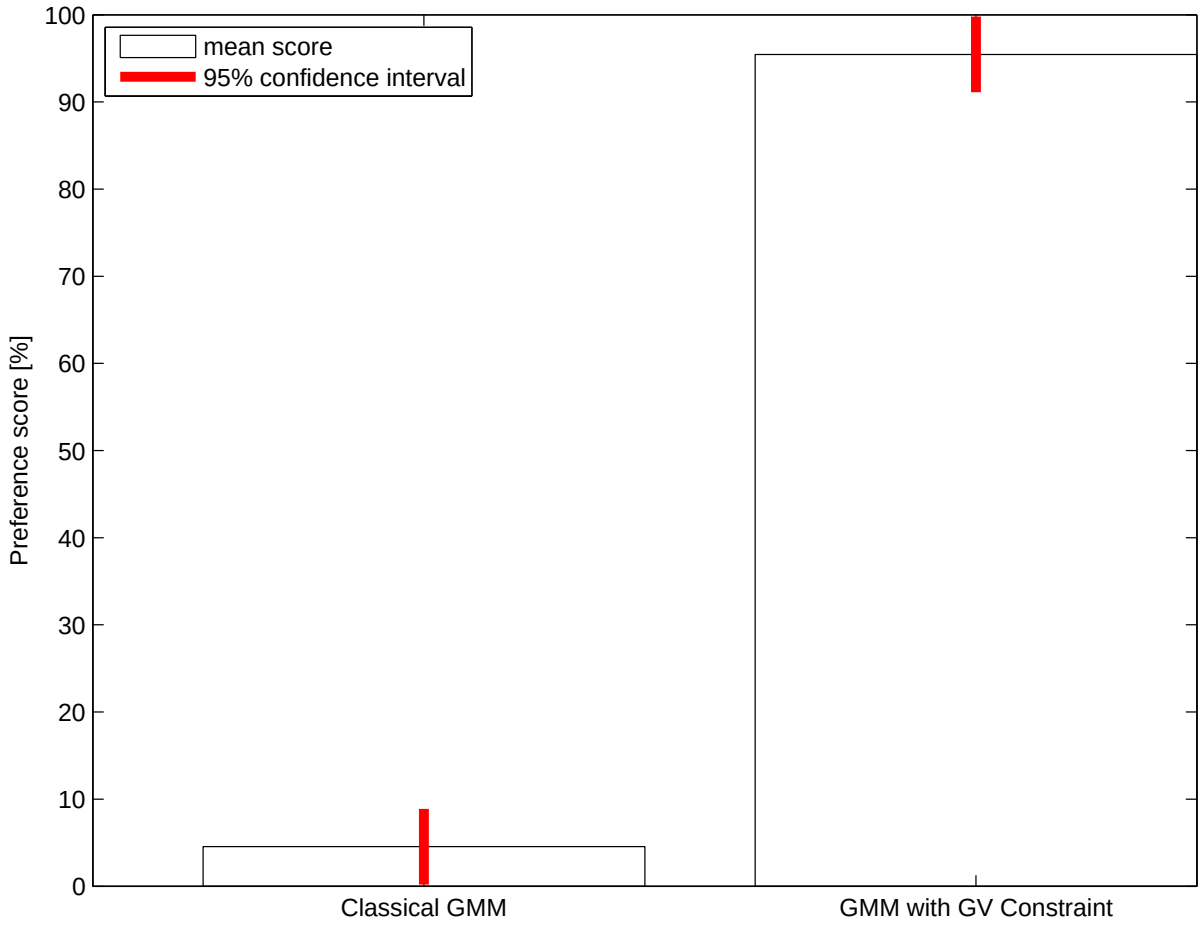
# Objective Measures - Results

| Conversion Method | Mean LSD [dB] | Mean Norm. GV |
|---|---|---|
| Original Source-Target | 8.6 | 1 |
| Classical GMM | 6.2 | 0.1 |
| Constrained GMM | 7.3 | 0.9 |

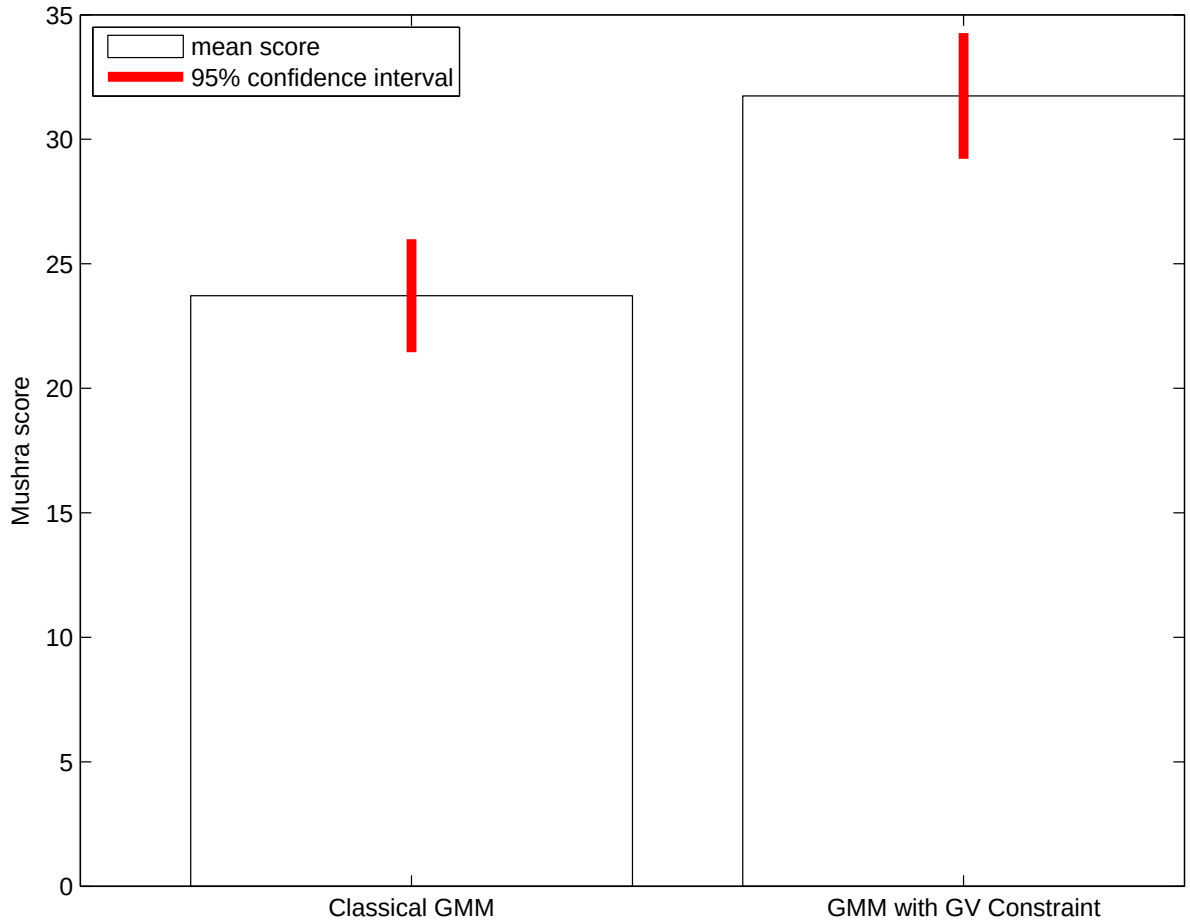
# Subjective Measures

- Quality -
  - Preference test - AB
  - Multi Stimulus test with Hidden Reference and Anchor (MUSHRA)
- Individuality (Similarity to the target) - XAB text
- 10 different sentences, 10 non-experts listeners
- Compared conversions -
  - The classical conversion
  - The proposed constrained conversion

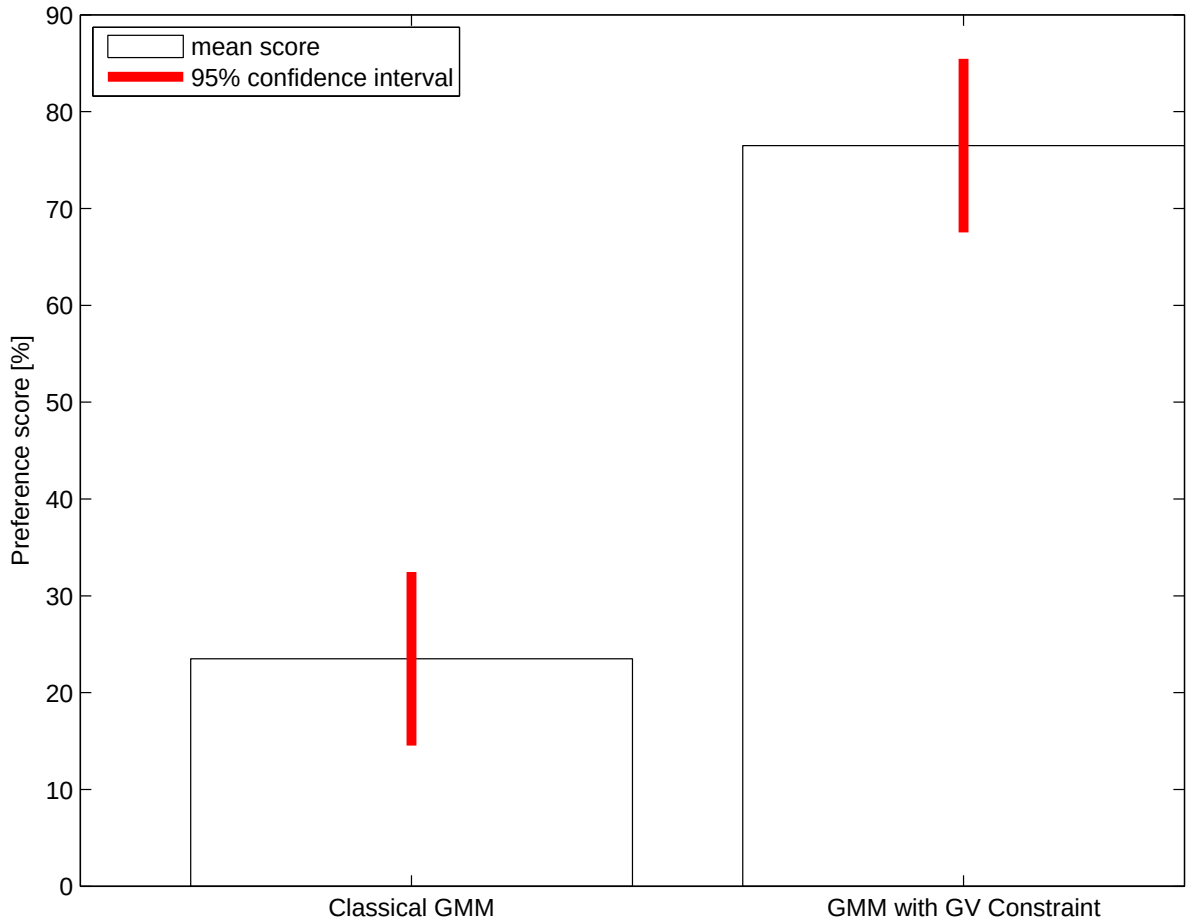
# Quality Preference Test - Results



# Mushra - Results



# Preference Individuality Test - Results



# Conclusion

- To deal with the muffling effect we propose a constrained formulation of the classical conversion:
  - Minimize the LSD under a GV constraint
- Objectively - GV is significantly increased, but also the mean LSD
- Subjectively (Quality & Similarity) - the constrained solution outperformed the classical solution
- **Further work** - GV enhancement in case of non-diagonal covariance GMM presents a major computational complexity challenge

*Thank You*

# Line Spectral Frequencies (LSF)

- An auto-regressive process can be represented using an all-pole filter - $\frac{1}{A(z)}$
- An alternative representation for $A(z)$ is using the normalized frequencies $w_i$, of the roots of the polynomials, defined by:

$$\begin{aligned} P(z) &= A(z)\left[1 + z^{-(1+p)}\frac{A\left(z^{-1}\right)}{A(z)}\right] \\ Q(z) &= A(z)\left[1 - z^{-(1+p)}\frac{A\left(z^{-1}\right)}{A(z)}\right] \end{aligned}$$

where $p$ is the order of $A(z)$.

# Mel Frequency Cepstrum Coefficients (MFCCs)

- The real Cepstrum coefficients $\{c_p\}_1^P$ relate to the spectral envelop $S(f)$ by:

$$log|S(f)| = c_0 + 2\sum_{p=1}^{P} c_p cos(2\pi fp)$$

- Define the vector holding the log of the harmonic amplitudes:

$$\mathbf{a} \equiv (log(A_1), log(A_2), ..., log(A_L))^T$$

- The real Cepstrum vector $\mathbf{c} \equiv (c_1, c_2, ..., c_P)^T$ can be evaluated by:

$$\mathbf{a} = \mathbf{M}^T\mathbf{c}$$

# Mel Frequency Cepstrum Coefficients (MFCCs) – Cont.

- where:

$$\mathbf{M}^T = \begin{pmatrix} 1 & 2cos\left(2\pi f_0\right) & \cdots & 2cos\left(2\pi f_0 \cdot P\right) \\ 1 & 2cos\left(2\pi 2f_0\right) & \cdots & 2cos\left(2\pi 2f_0 \cdot P\right) \\ \vdots & \vdots & \cdots & \vdots \\ 1 & 2cos\left(2\pi Lf_0\right) & \cdots & 2cos\left(2\pi Lf_0 \cdot P\right) \end{pmatrix}$$

- Usually $P < L$, so $\mathbf{c}$ is evaluated using LS estimation:

$$\mathbf{c} = \left(\mathbf{M}^T\mathbf{M}\right)^{-1}\mathbf{M}^T\mathbf{a}$$

# GMM Training

- The **source** training vectors, $\{\mathbf{x}^q\}_{q=1}^{Q}$, are divided into $M$ classes
- The vectors in every class are assumed to be jointly Gaussian:

$$p\left(\mathbf{x}^q\right) = \sum_{m=1}^{M} p\left(w_m\right) N\left(\mathbf{x}^q; \mu^{(\mathbf{x}),m}, \mathbf{\Sigma}^{(\mathbf{xx}),m}\right)$$
$$q = 1, ..., Q$$

$p\left(w_m\right)$ – the probability of the class $w_m$
$N\left(\cdot; \mu^{(\mathbf{x}),m}, \mathbf{\Sigma}^{(\mathbf{xx}),m}\right)$ – a normal distribution

- The GMM parameters, $\left\{p\left(w_m\right), \mu^{(\mathbf{x}),m}, \mathbf{\Sigma}^{(\mathbf{xx}),m}\right\}_{m=1}^{M}$, are usually evaluated using the Expectation Maximization (EM) procedure, [Dempster et. al, 1977].

# Full Vs. Diagonal GMM

- Full GMM conversion
  - The trained matrices $\{\mathbf{\Sigma}^{(\mathbf{xx}),\mathbf{m}}\}_{\mathbf{1}}^{\mathbf{M}}$ are full
  - Requires a large data set
- Diagonal GMM Conversion
  - The elements of the spectral feature vectors are assumed to be uncorrelated, so the trained covariance matrices are diagonal:

$$\{\mathbf{\Sigma}^{(\mathbf{xx}),m}\}_{p,p} = \left(\sigma_p^{(\mathbf{xx}),m}\right)^2$$

  - Requires a smaller training set
- In practice, training a full conversion is often replaced by a diagonal conversion with more Gaussians (larger $M$)

# A Diagonal Conversion

- A linear conversion function is defined using the trained GMM parameters:

$$\mathcal{F}_p^{(LS-GMM)}\{x_p\} = \sum_{m=1}^{M} p\left(w_m|x^p\right)\left(\nu_p{}^m + \frac{\gamma_p^m}{\sigma_p{}^{(xx),m}}\left(x_p - \mu_p{}^{(x),m}\right)\right)$$

- Where $\{\nu_p{}^m, \gamma_p^m\}_{m=1}^{M}, \quad p = 1, \ldots P$ are the conversion parameters

# Training a Diagonal GMM-Based Conversion

## LS estimation of the conversion parameters:

$$\mathbf{y}^p = \begin{pmatrix} \mathbf{P} & \vdots & \mathbf{D}^p \end{pmatrix} \cdot \begin{pmatrix} \mathbf{v}^p \\ \cdots \\ \mathbf{g}^p \end{pmatrix}$$

- $\mathbf{y}^p$ – the $p$-th elements of the target training vectors
- $\{\mathbf{P}\}_{m,q} \triangleq p\left(w_m | x_p^q\right)$
- $\{\mathbf{D}^p\}_{q,m} \triangleq p\left(w_m | x_p^q\right) \frac{\left(x_p^q - \mu_p^{(x),m}\right)}{\sigma^{q,m}}$
- $\mathbf{v}^p \triangleq \begin{pmatrix} \nu_p{}^1 & \cdots & \nu_p{}^M \end{pmatrix}^T$
- $\mathbf{g}^p \triangleq \begin{pmatrix} \gamma_p^1 & \cdots & \gamma_p^M \end{pmatrix}^T$
- $q = 1, ..., Q, \quad m = 1, ...M$

# Unconstrained Formulation

## LS estimation of the conversion parameters:

$$\mathbf{y}^p = \begin{pmatrix} \mathbf{P} & \vdots & \mathbf{D}^p \end{pmatrix} \cdot \begin{pmatrix} \mathbf{v}^p \\ \cdots \\ \mathbf{g}^p \end{pmatrix} \quad \Rightarrow \quad \min_{\mathbf{q}^p} \|\mathbf{A}^p\mathbf{q}^p - \mathbf{b}^p\|^2$$

| Notation | Definition | Dimmension |
|---|---|---|
| $\mathbf{A}^p$ | $\begin{pmatrix} \mathbf{P} & \vdots & \mathbf{D}^p \end{pmatrix}$ | $Q \times 2M$ |
| $\mathbf{q}^p$ | $\begin{pmatrix} \mathbf{v}^p \\ \cdots \\ \mathbf{g}^p \end{pmatrix}$ | $2M \times 1$ |
| $\mathbf{b}^p$ | $\mathbf{y}^p$ | $Q \times 1$ |