“Voice Conversion Based on Maximum Likelihood Estimation of Spectral Parameter Trajectory”

- Tomoki Toda, Alan W. Black and Keiichi Tokuda

Under the guidance of:

Dr. R. hegde

By:-

Mayank Sirotiya (Y8104036)

Vipul Arora (Y5508)

Abstract:

- This paper describes a method for Spectral Estimation for Voice Conversion application, based on Maximum Likelihood Estimation technique.
- It also presents the conventional methods of spectral parameter conversion and addresses the problems with them.

Introduction:

- Voice conversion technology enables a user to transform one person's speech pattern into another pattern with distinct characteristics

- A mapping function is used which consists of utterance pairs of source and target voices

Applications

- Speaker conversion
- Cross-language speaker conversion
- Narrow-band to wide-band speech for telecommunication
- Speaking aid
- Modeling of speech production etc.

Classical Approaches and their Limitations:

A. Codebook mapping based on hard clustering and discrete mapping

$$\hat{\boldsymbol{y}}_t = \boldsymbol{c}_m^{(y)}$$

B. Fuzzy vector quantization, for soft clustering

$$\hat{\boldsymbol{y}}_t = \sum_{m=1}^{M} w_{m,t}^{(x)} \boldsymbol{c}_m^{(y)}$$

Classical Approaches and their Limitations:

C. More variable representation, by modeling a difference vector

$$\hat{\boldsymbol{y}}_t = \boldsymbol{x}_t + \sum_{m=1}^{M} w_{m,t}^{(x)} \left(\boldsymbol{c}_m^{(y)} - \boldsymbol{c}_m^{(x)} \right)$$

D. Method using linear multivariate regression (LMR)

$$\hat{\boldsymbol{y}}_t = \boldsymbol{A}_m \boldsymbol{x}_t + \boldsymbol{b}_m$$

Classical Approaches and their Limitations:

E. Gaussian mixture model

It realizes continuous mapping based on soft clustering

$$\hat{\boldsymbol{y}}_t = \sum_{m=1}^{M} w_{m,t}^{(x)} \left(\boldsymbol{A}_m \boldsymbol{x}_t + \boldsymbol{b}_m \right)$$

Conventional GMM based mapping

- The joint probability density of the source and target feature vectors is

$$P\left(z_t|\lambda^{(z)}\right) = \sum_{m=1}^{M} w_m N\left(z_t; \mu_m^{(Z)}, \Sigma_m^{(z)}\right)$$

where z_t is a joint vector $\left[\mathbf{x}_t^T, \mathbf{y}_t^T\right]^T$
and the mean vector and covariance matrix are written as

$$\mu_m^{(Z)} = \begin{bmatrix} \mu_m^{(x)} \\ \mu_m^{(y)} \end{bmatrix}, \Sigma_m^{(z)} = \begin{bmatrix} \Sigma_m^{(xx)} & \Sigma_m^{(xy)} \\ \Sigma_m^{(yx)} & \Sigma_m^{(yy)} \end{bmatrix}$$

Conventional GMM based mapping

- Conditional probability density can also be represented as

$$P\left(\mathbf{y}_t|\mathbf{x}_t,\lambda^{(z)}\right) = \sum_{m=1}^{M} P\left(m|\mathbf{x}_t,\lambda^{(z)}\right) P\left(\mathbf{y}_t|\mathbf{x}_t,m,\lambda^{(z)}\right)$$

where

$$P\left(m|\mathbf{x}_t,\lambda^{(z)}\right) = \frac{w_m N\left(\mathbf{x}_t;\mu_m^{(x)},\mathbf{\Sigma}_m^{(xx)}\right)}{\sum_{m=1}^{M} w_n N\left(\mathbf{x}_t;\mu_n^{(x)},\mathbf{\Sigma}_n^{(xx)}\right)}$$

$$P\left(\mathbf{y}_t|\mathbf{x}_t,m,\lambda^{(z)}\right) = N\left(\mathbf{y}_t;\mathrm{E}_{m,t}^{(y)},\mathrm{D}_m^{(y)}\right)$$

Conventional GMM based mapping

- The mean vector and the covariance matrix of m'th conditional probability distribution are written as

$$\mathrm{E}_{m,t}^{(y)} = \mu_m^{(y)} + \Sigma_m^{(yx)} \Sigma_m^{(xx)-1} \left(\mathrm{x}_t - \mu_m^{(x)} \right)$$

$$\mathrm{D}_m^{(y)} = \Sigma_m^{(yy)} - \Sigma_m^{(yx)} \Sigma_m^{(xx)-1} \Sigma_m^{(xy)}$$

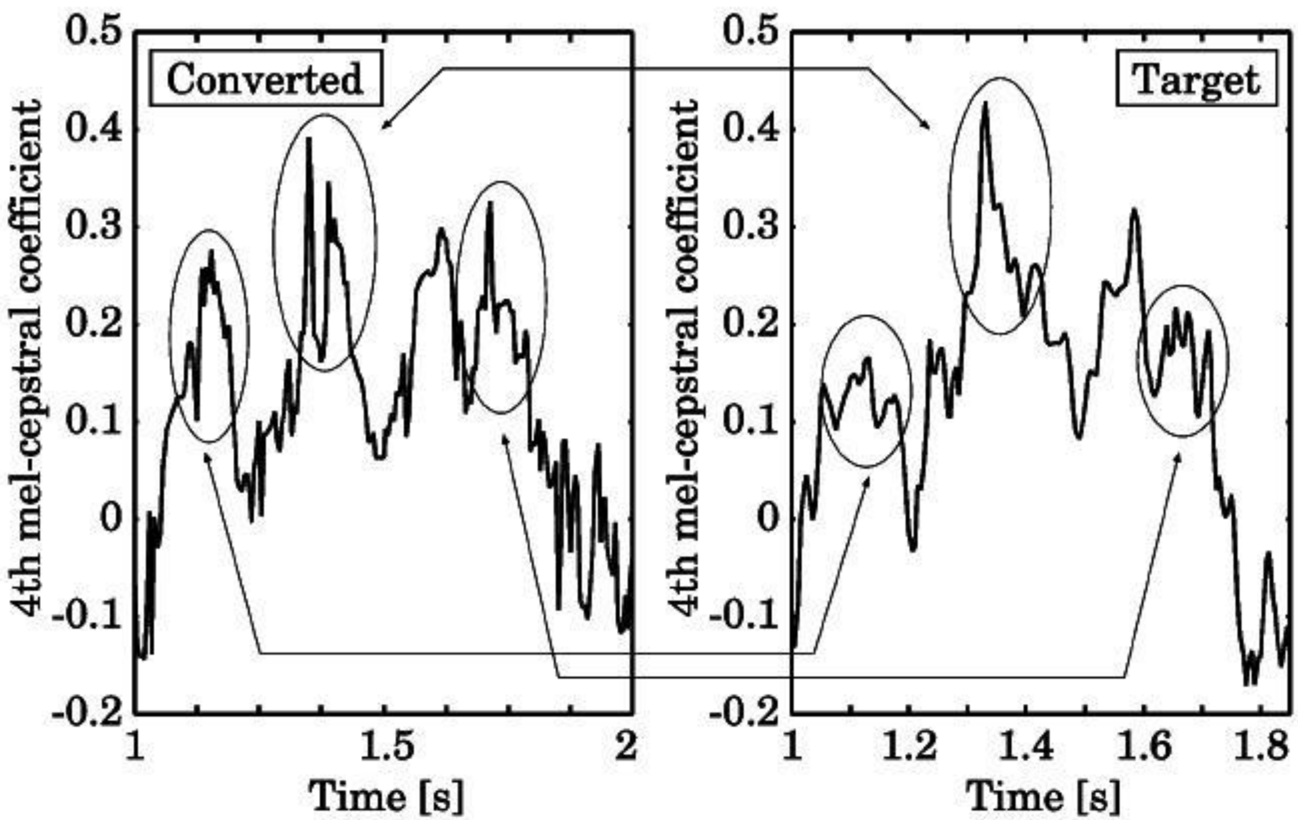
Conventional GMM based mapping

- In conventional method the conversion is based on MMSE as follows

$$
\begin{aligned}
\hat{\boldsymbol{y}}_t &= E[\boldsymbol{y}_t | \boldsymbol{x}_t] \\
&= \int P\left(\boldsymbol{y}_t | \boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}\right) \boldsymbol{y}_t d\boldsymbol{y}_t \\
&= \int \sum_{m=1}^{M} P\left(m | \boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}\right) P\left(\boldsymbol{y}_t | \boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(z)}\right) \boldsymbol{y}_t d\boldsymbol{y}_t \\
&= \sum_{m=1}^{M} P\left(m | \boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}\right) \boldsymbol{E}_{m,t}^{(y)}
\end{aligned}
$$

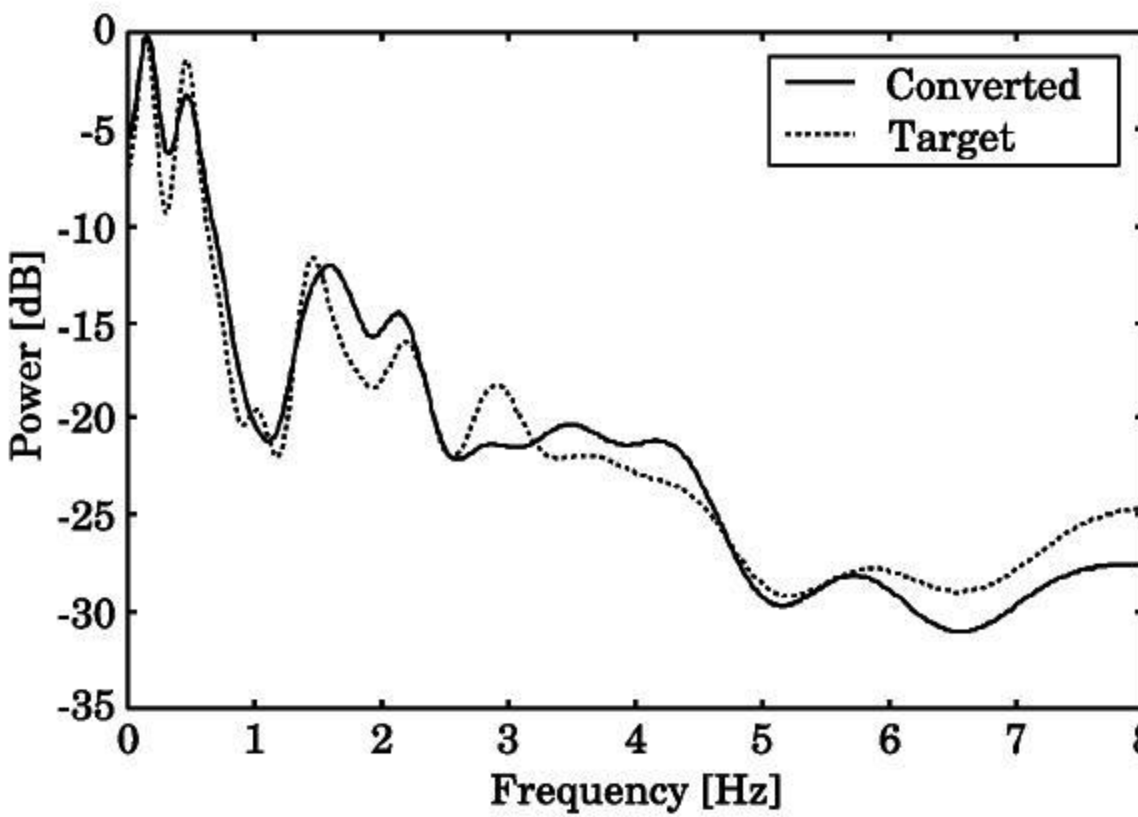
Drawbacks

1. Time-independent mapping



Drawbacks

2. Oversmoothing



PROPOSED SPECTRAL CONVERSION

- Trajectory based spectral conversion process, instead of conventional frame based one.

$$\hat{\boldsymbol{y}} = f(\boldsymbol{x})$$

where,

$$\boldsymbol{x} = \left[\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \ldots, \boldsymbol{x}_t^\top, \ldots, \boldsymbol{x}_T^\top\right]^\top$$
$$\boldsymbol{y} = \left[\boldsymbol{y}_1^\top, \boldsymbol{y}_2^\top, \ldots, \boldsymbol{y}_t^\top, \ldots, \boldsymbol{y}_T^\top\right]^\top$$

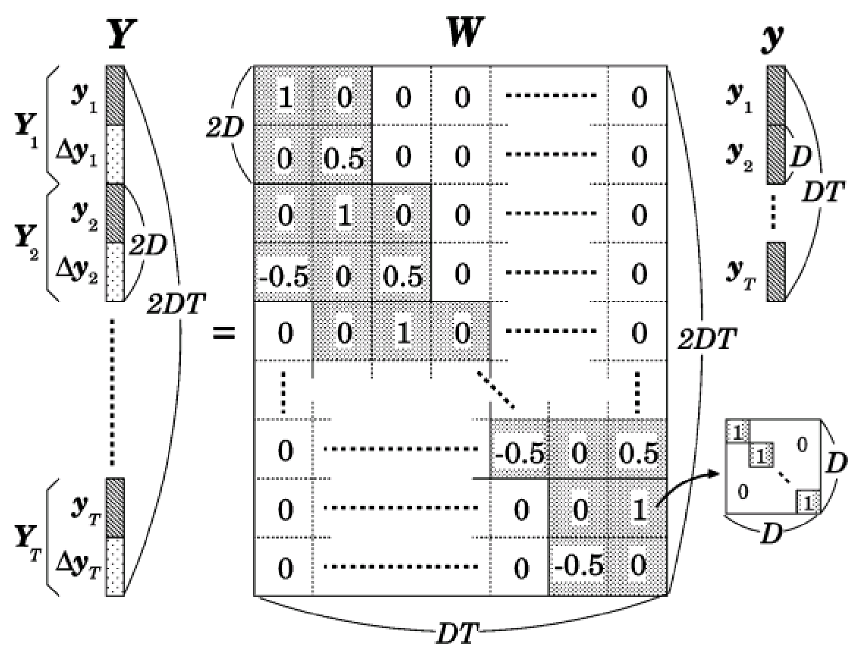
- Conversion considering feature correlation between frames (dynamic features)

$$X = \left[X_1^\top, X_2^\top, \ldots, X_t^\top, \ldots, X_T^\top\right]^\top$$

$$Y = \left[Y_1^\top, Y_2^\top, \ldots, Y_t^\top, \ldots, Y_T^\top\right]^\top$$

$$X_t = \left[x_t^\top, \Delta x_t^\top\right]^\top \text{ and } Y_t = \left[y_t^\top, \Delta y_t^\top\right]^\top$$

$$Y = Wy$$



MLE of Parameter Trajectory

- The joint vector is

$$Z_t = \left[X_t^\top, Y_t^\top\right]^\top$$

- The GMM of joint probability density

$$P(\boldsymbol{Z}_t|\boldsymbol{\lambda}^{(Z)})$$

  is trained using conventional training framework.

- Likelihood Function to be maximized:

$$P\left(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\lambda}^{(Z)}\right) = \prod_{t=1}^{T}\sum_{m=1}^{M} P\left(m|\boldsymbol{X}_t,\boldsymbol{\lambda}^{(Z)}\right) \times P\left(\boldsymbol{Y}_t|\boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(Z)}\right)$$

where,

$$P\left(m|\boldsymbol{X}_t, \boldsymbol{\lambda}^{(Z)}\right) = \frac{w_m \mathcal{N}\left(\boldsymbol{X}_t; \boldsymbol{\mu}_m^{(X)}, \boldsymbol{\Sigma}_m^{(XX)}\right)}{\sum_{n=1}^{M} w_n \mathcal{N}\left(\boldsymbol{X}_t; \boldsymbol{\mu}_n^{(X)}, \boldsymbol{\Sigma}_n^{(XX)}\right)}$$

$$P\left(\boldsymbol{Y}_t|\boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(Z)}\right) = \mathcal{N}\left(\boldsymbol{Y}_t; \boldsymbol{E}_{m,t}^{(Y)}, \boldsymbol{D}_m^{(Y)}\right)$$

here,

$$\boldsymbol{E}_{m,t}^{(Y)} = \boldsymbol{\mu}_m^{(Y)} + \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(XX)^{-1}} \left(\boldsymbol{X}_t - \boldsymbol{\mu}_m^{(X)}\right)$$

$$\boldsymbol{D}_m^{(Y)} = \boldsymbol{\Sigma}_m^{(YY)} - \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(XX)^{-1}} \boldsymbol{\Sigma}_m^{(XY)}$$

Derivation of Conditional Probability

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$$

$$\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1} \qquad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

$$\begin{aligned} -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \\ -\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Lambda}_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b) \\ -\frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b-\boldsymbol{\mu}_b) \end{aligned}$$

Derivation of Conditional Probability

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \text{const}$$

- Second order term is,

$$-\frac{1}{2}\mathbf{x}_a^{\mathrm{T}}\boldsymbol{\Lambda}_{aa}\mathbf{x}_a$$

- First order term is,

$$\mathbf{x}_a^{\mathrm{T}}\left\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\right\}$$

i) EM Algorithm

$$\hat{\boldsymbol{y}} = \arg\max P\left(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(Z)}\right)$$

$$\begin{aligned} Q(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) &= \sum_{\text{all } \boldsymbol{m}} P\left(\boldsymbol{m}|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\lambda}^{(Z)}\right) \log P\left(\hat{\boldsymbol{Y}}, \boldsymbol{m}|\boldsymbol{X}, \boldsymbol{\lambda}^{(Z)}\right) \\ &= \sum_{t=1}^{T}\sum_{m=1}^{M} P\left(m|\boldsymbol{X}_t, \boldsymbol{Y}_t, \boldsymbol{\lambda}^{(Z)}\right) \log P\left(\hat{\boldsymbol{Y}}_t, m|\boldsymbol{X}_t, \boldsymbol{\lambda}^{(Z)}\right) \\ &= \sum_{t=1}^{T}\sum_{m=1}^{M} \gamma_{m,t}\left(-\frac{1}{2}\hat{\boldsymbol{Y}}_t^{\top}\boldsymbol{D}_m^{(Y)^{-1}}\hat{\boldsymbol{Y}}_t + \hat{\boldsymbol{Y}}_t^{\top}\boldsymbol{D}_m^{(Y)^{-1}}\boldsymbol{E}_{m,t}^{(Y)}\right) + \overline{K} \end{aligned}$$

$$
\begin{aligned}
Q(\boldsymbol{Y},\hat{\boldsymbol{Y}}) &= \sum_{t=1}^{T} -\frac{1}{2}\hat{\boldsymbol{Y}}_t^\top \overline{\boldsymbol{D}_t^{(Y)^{-1}}}\hat{\boldsymbol{Y}}_t + \hat{\boldsymbol{Y}}_t^\top \overline{\boldsymbol{D}_t^{(Y)^{-1}}\boldsymbol{E}_t^{(Y)}} + \overline{K} \\
&= -\frac{1}{2}\hat{\boldsymbol{Y}}^\top \overline{\boldsymbol{D}^{(Y)^{-1}}}\hat{\boldsymbol{Y}} + \hat{\boldsymbol{Y}}^\top \overline{\boldsymbol{D}^{(Y)^{-1}}\boldsymbol{E}^{(Y)}} + \overline{K} \\
&= -\frac{1}{2}\hat{\boldsymbol{y}}^\top \boldsymbol{W}^\top \overline{\boldsymbol{D}^{(Y)^{-1}}}\boldsymbol{W}\hat{\boldsymbol{y}} + \hat{\boldsymbol{y}}^\top \boldsymbol{W}^\top \overline{\boldsymbol{D}^{(Y)^{-1}}\boldsymbol{E}^{(Y)}} + \overline{K}
\end{aligned}
$$

$$
\frac{\partial Q(\boldsymbol{Y},\hat{\boldsymbol{Y}})}{\partial \boldsymbol{y}} = -\boldsymbol{W}^\top \overline{\boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}}}\boldsymbol{W}\boldsymbol{y} + \boldsymbol{W}^\top \overline{\boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}}\boldsymbol{E}_{\hat{\boldsymbol{m}}}^{(Y)}}
$$

- Equating it to zero, we get

$$
\hat{\boldsymbol{y}} = \left(\boldsymbol{W}^\top \overline{\boldsymbol{D}^{(Y)^{-1}}}\boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \overline{\boldsymbol{D}^{(Y)^{-1}}\boldsymbol{E}^{(Y)}}
$$

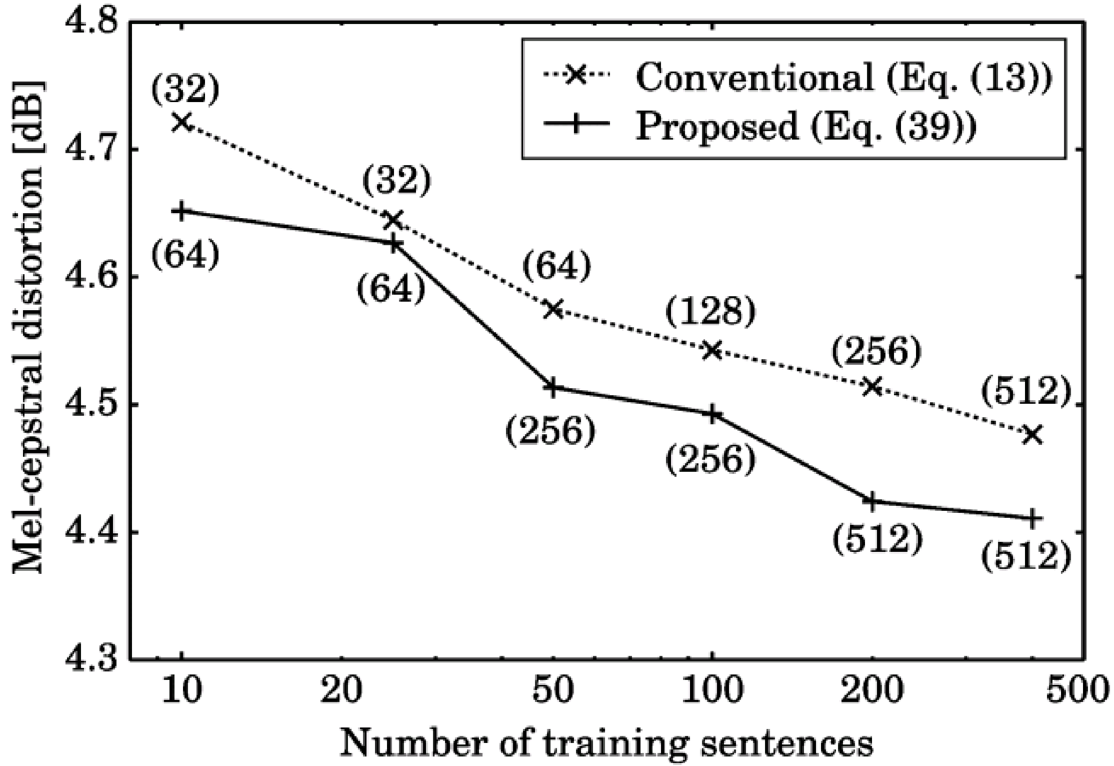
ii) Approximation with suboptimum mixture sequence

$$\hat{\boldsymbol{m}} = \operatorname{argmax} P\left(\boldsymbol{m} | \boldsymbol{X}, \boldsymbol{\lambda}^{(Z)}\right)$$

$$\mathcal{L} = \log P\left(\hat{\boldsymbol{m}} | \boldsymbol{X}, \boldsymbol{\lambda}^{(Z)}\right) P\left(\boldsymbol{Y} | \boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}^{(Z)}\right)$$

$$\hat{\boldsymbol{y}} = \left(\boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}} \boldsymbol{E}_{\hat{\boldsymbol{m}}}^{(Y)}$$

Results:



(Mel Cepstral distortion before conversion is 7.30 dB)

Summary

- GMM based Feature mapping
- Conditional Gaussian Distributions
- Maximum Likelihood technique for GMMs
- Expectation Maximization Algorithm