# Vector Taylor Series Expansion with Auditory Masking for Noise Robust Speech Recognition

**Biswajit Das** and Ashish Panda

TCS Innovation Labs - Mumbai, India

*b.das@tcs.com*

20 Oct, 2016

# Outlines

- Introduction
- Proposed method
- Algorithm
- Experimental Setup
- Experimental Results
- Conclusion

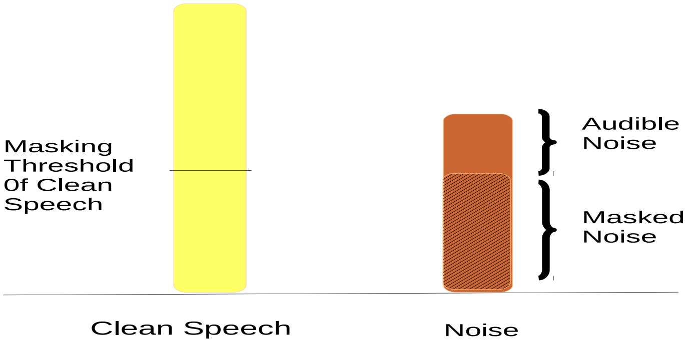
# Introduction

- Existing Methods
  - Vector Taylor Series (VTS) expansion for Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) (J. Li et al., 2007)
  - Psychoacoustic Compensation (Psy-Comp) technique for GMM-HMM model (B Das and A Panda, 2015).
  - VTS technique for feature enhancement for Deep Neural Network (DNN) (B Li, 2013).

# Introduction

- Existing Methods
  - Vector Taylor Series (VTS) expansion for Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) (J. Li et al., 2007)
  - Psychoacoustic Compensation (Psy-Comp) technique for GMM-HMM model (B Das and A Panda, 2015).
  - VTS technique for feature enhancement for Deep Neural Network (DNN) (B Li, 2013).
- Proposed method
  - Incorporation of auditory masking into the Vector Taylor Series for corrupting clean model parameters.
  - Use of Minimum Mean Square Error (MMSE) to extract clean features.

# Speech Masking Noise



# Traditional vector Taylor series

- Degraded speech model in spectral domain

$$Y = XH + N, \tag{1}$$

where $Y$, $X$, $H$ and $N$ are degraded speech, clean speech, channel factor, and additive noise respectively in spectral domain.

# Traditional vector Taylor series

- Degraded speech model in spectral domain

$$Y = XH + N, \tag{1}$$

where $Y$, $X$, $H$ and $N$ are degraded speech, clean speech, channel factor, and additive noise respectively in spectral domain.

- VTS Corruption function

$$\vec{y^s} = \vec{x^s} + \vec{h^s} + Clog(1 + exp(C^{-1}(\vec{n^s} - \vec{x^s} - \vec{h^s}))), \tag{2}$$

where $s$ indicates the static part of all variables. $\vec{y}$, $\vec{x}$, $\vec{h}$ and $\vec{n}$ are distorted speech, clean speech, channel factor and additive noise respectively. $C$ and $C^{-1}$ are the discrete cosine transform matrix and its inverse respectively.

# Proposed Method

- Proposed degraded speech model in spectral domain

$$Y_f = W_f X_f H_f + N_f \tag{3}$$

where $H_f$ is the channel factor and $N_f$ is the additive noise. The masking factor $W_f$ can be defined as follows:

$$W_f = \frac{X_f - 10^{\frac{T_{mf}}{20}}}{X_f}. \tag{4}$$

$T_{mf}$ is the masking threshold of the clean speech $X_f$.

# Proposed Method

- Proposed degraded speech model in spectral domain

$$Y_f = W_f X_f H_f + N_f \tag{3}$$

where $H_f$ is the channel factor and $N_f$ is the additive noise. The masking factor $W_f$ can be defined as follows:

$$W_f = \frac{X_f - 10^{\frac{T_{mf}}{20}}}{X_f}. \tag{4}$$

$T_{mf}$ is the masking threshold of the clean speech $X_f$.

- Masking threshold is calculated as follows:

$$T_{xf} = 20\log_{10}(\mu_{xf}) - 0.275.h_f - 6.025 \quad \mathrm{(dB)} \tag{5}$$

where $h_f$ is central frequency of mel-filter in Bark scale.

- Proposed Corruption function

$$\vec{y^s} = \vec{x^s} + \vec{h^s} + \vec{w^s} + Clog(1 + exp(C^{-1}(\vec{n^s} - \vec{x^s} - \vec{h^s} - \vec{w^s}))), \quad (6)$$

where $\vec{y}$, $\vec{x}$, $\vec{h}$, $\vec{w}$, and $\vec{n}$ are distorted speech, clean speech, channel factor, masking factor and additive noise respectively and all these parameters are in MFCC domain.

# Proposed Method

- Proposed Corruption function

$$\vec{y^s} = \vec{x^s} + \vec{h^s} + \vec{w^s} + Clog(1 + exp(C^{-1}(\vec{n^s} - \vec{x^s} - \vec{h^s} - \vec{w^s}))), \quad (6)$$

  where $\vec{y}$, $\vec{x}$, $\vec{h}$, $\vec{w}$, and $\vec{n}$ are distorted speech, clean speech, channel factor, masking factor and additive noise respectively and all these parameters are in MFCC domain.
- The Jacobian of the mismatch function with respect to clean speech parameter

$$G = C \bullet diag\left(\frac{1}{1 + exp(C^{-1}(\vec{\mu_n} - \vec{\mu_x} - \vec{w} - \vec{h}))}\right) \bullet C^{-1}. \quad (7)$$

# Proposed Method

- Proposed Corruption function

$$\vec{y^s} = \vec{x^s} + \vec{h^s} + \vec{w^s} + Clog(1 + exp(C^{-1}(\vec{n^s} - \vec{x^s} - \vec{h^s} - \vec{w^s}))), \quad (6)$$

where $\vec{y}$, $\vec{x}$, $\vec{h}$, $\vec{w}$, and $\vec{n}$ are distorted speech, clean speech, channel factor, masking factor and additive noise respectively and all these parameters are in MFCC domain.

- The Jacobian of the mismatch function with respect to clean speech parameter

$$G = C \bullet diag\left(\frac{1}{1 + exp(C^{-1}(\vec{\mu_n} - \vec{\mu_x} - \vec{w} - \vec{h}))}\right) \bullet C^{-1}. \quad (7)$$

- The model mean corruption

$$\vec{\mu}_y = \vec{\mu}_x + \vec{h} + \vec{w} + Clog(1 + exp(C^{-1}(\vec{\mu_n} - \vec{\mu}_x - \vec{w} - \vec{h}))) \quad (8)$$

# Proposed Method

- Proposed Corruption function

$$\vec{y^s} = \vec{x^s} + \vec{h^s} + \vec{w^s} + Clog(1 + exp(C^{-1}(\vec{n^s} - \vec{x^s} - \vec{h^s} - \vec{w^s}))), \quad (6)$$

  where $\vec{y}$, $\vec{x}$, $\vec{h}$, $\vec{w}$, and $\vec{n}$ are distorted speech, clean speech, channel factor, masking factor and additive noise respectively and all these parameters are in MFCC domain.
- The Jacobian of the mismatch function with respect to clean speech parameter

$$G = C \bullet diag\left(\frac{1}{1 + exp(C^{-1}(\vec{\mu_n} - \vec{\mu_x} - \vec{w} - \vec{h}))}\right) \bullet C^{-1}. \quad (7)$$

- The model mean corruption

$$\vec{\mu}_y = \vec{\mu}_x + \vec{h} + \vec{w} + Clog(1 + exp(C^{-1}(\vec{\mu_n} - \vec{\mu}_x - \vec{w} - \vec{h}))) \quad (8)$$

- The model variance corruption

$$\Sigma_y \approx G\Sigma_x G^T + (I - G)\Sigma_n(I - G)^T. \quad (9)$$

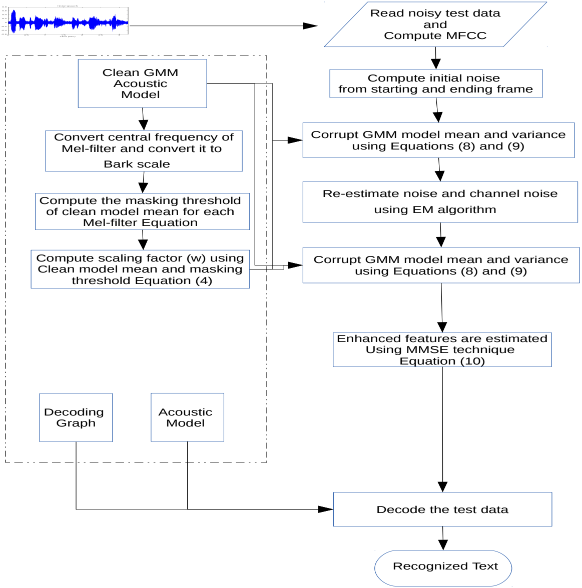
# Proposed Method

- We need a GMM model, trained on clean speech utterances
- Estimation of Enhanced Features using MMSE technique

$$\begin{aligned} \vec{x}_{MMSE} &= E(\vec{x}|\vec{o}) = \int \vec{x} p(\vec{x}|\vec{o}) dx \\ &= \vec{o} - \sum_{m=0}^{M-1} p(\vec{o}|\lambda_{ym})(\vec{\mu}_{ym} - \vec{\mu}_{xm}), \end{aligned} \tag{10}$$

  - $\vec{o}$ : noisy speech features
  - $p(\vec{o}|\lambda_{ym})$ : posterior probability for the $m^{th}$ Gaussian mixture component of the noise compensated GMM.
  - $\vec{\mu}_{ym}$ : $m^{th}$ component of the noise compensated model mean.
  - $\vec{\mu}_{xm}$ : $m^{th}$ component of the clean model mean.

# Algorithm for Model Compensation



# Experimental Setup

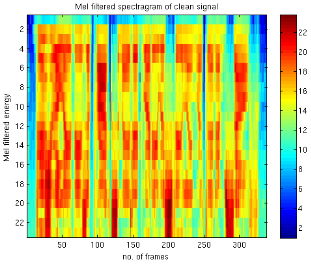
- Speech Corpus: TIMIT
- ASR Toolkit: Kaldi ASR toolkit
- Feature: Mel Filter Cepstral Coefficient (MFCC).
- Model Description (DNN-HMM):
    - Number of hidden layer : 2
- Adding noise: Used Filtering and Noise Adding Tool (FaNT)
- Clean train data and noise corrupted test data.
- Noise Type: Babble, Hfchannel, F-16 from NOISEX-92 database and Street noise (We collected)
- SNR Level: 0dB, 5dB, 10dB and 15dB

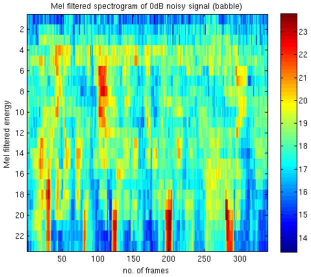
# Different Systems

- Different acoustic model:
  1. TRI1: It has been obtained after standard GMM-HMM approach. It is a triphone model.
  2. TRI2: Applied Linear Discriminative Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) at the time of training.
  3. TRI3: Speaker Adaptive Training (SAT) along with LDA and MLLT for speaker dependent acoustic model.
  4. DNN: DNN architecture is used instead of GMM for acoustic modeling.

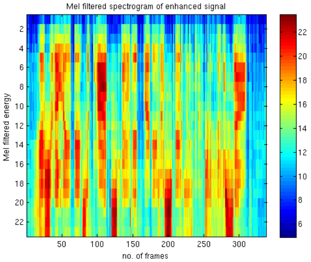
# Different Systems

- Different acoustic model:
  1. TRI1: It has been obtained after standard GMM-HMM approach. It is a triphone model.
  2. TRI2: Applied Linear Discriminative Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) at the time of training.
  3. TRI3: Speaker Adaptive Training (SAT) along with LDA and MLLT for speaker dependent acoustic model.
  4. DNN: DNN architecture is used instead of GMM for acoustic modeling.
- Different features enhancement techniques:
  1. Baseline: No enhancement technique.
  2. VTS: We have enhanced feature using traditional VTS method.
  3. Proposed method : We have introduced masking effect into VTS method to enhance features.




1. Spectrogram of clean signal
2. Spectrogram of 0dB noisy signal (Babble) .
3. Enhanced spectrogram with proposed method.



# Experimental Results

Table: Phoneme error rate for various methods for "hfchannel" with different noise level

| | HFCHANNEL | | | | | | | | | | | | Average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0DB | | | 5DB | | | 10DB | | | 15DB | | | | | |
| | Baseline | VTS | Proposed | Baseline | VTS | Proposed | Baseline | VTS | Proposed | Baseline | VTS | Proposed | Baseline | VTS | Proposed |
| TRI1 | 81.5 | 66.1 | 63.7 | 72.3 | 61.6 | 53.9 | 62.9 | 45.9 | 44.5 | 50.3 | 41.3 | 37.3 | 66.75 | 53.73 | 49.85 |
| TRI2 | 83.8 | 65.2 | 64.3 | 72.9 | 61.5 | 53.8 | 60.5 | 45.5 | 43.3 | 46.7 | 41.5 | 36 | 65.98 | 53.43 | 49.35 |
| TRI3 | 81.6 | 64.6 | 63.6 | 72.6 | 61.3 | 54.1 | 60.1 | 45.3 | 44.5 | 46.4 | 39.4 | 36.4 | 65.18 | 52.65 | 49.65 |
| DNNs | 79.6 | 60.5 | 59.3 | 64.2 | 56 | 49.8 | 47.6 | 41.9 | 40.8 | 36.3 | 37.3 | 33.1 | 56.93 | 48.93 | 45.75 |

Table: Phoneme error rate for various methods for "f-16" with different noise level

| | F-16 | | | | | | | | | | | | Average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0DB | | | 5DB | | | 10DB | | | 15DB | | | | | |
| | Baseline | VTS | Proposed | Baseline | VTS | Proposed | Baseline | VTS | Proposed | Baseline | VTS | Proposed | Baseline | VTS | Proposed |
| TRI1 | 87.8 | 72.3 | 69.2 | 78.3 | 61.6 | 58.6 | 66.9 | 50.8 | 47.9 | 53.5 | 41.3 | 38.9 | 71.63 | 56.5 | 53.65 |
| TRI2 | 89.2 | 71 | 68.7 | 82.4 | 61.5 | 58.9 | 70.5 | 50.9 | 49.6 | 55.6 | 41.5 | 38.9 | 74.43 | 56.23 | 54.03 |
| TRI3 | 90.5 | 71.2 | 68.2 | 81 | 61.3 | 59.5 | 70 | 51.6 | 49.6 | 55.1 | 42 | 39.6 | 74.15 | 56.53 | 54.23 |
| DNNs | 89.5 | 67.4 | 64.7 | 78.2 | 56 | 54.3 | 58.5 | 46.9 | 44.2 | 42.1 | 37.3 | 35.3 | 67.07 | 51.9 | 49.62 |

# Experimental Results ..

Table: Phoneme error rate for various methods for "babble" with different noise level

| | BABBLE | | | | | | | | | | | | Average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0DB | | | 5DB | | | 10DB | | | 15DB | | | | | |
| | Baseline | VTS | Proposed | Baseline | VTS | Proposed | Baseline | VTS | Proposed | Baseline | VTS | Proposed | Baseline | VTS | Proposed |
| TRI1 | 79.5 | 72.3 | 69.8 | 69.5 | 58.3 | 57 | 57.2 | 47.6 | 46.6 | 46.1 | 40 | 38.4 | 63.08 | 54.55 | 52.95 |
| TRI2 | 82.6 | 72.8 | 70.2 | 75 | 59.4 | 58 | 62 | 48.6 | 47.8 | 48.2 | 41.3 | 39.8 | 66.95 | 55.53 | 53.95 |
| TRI3 | 81.7 | 72.9 | 70.3 | 73.2 | 60.3 | 59.1 | 61.2 | 49.7 | 48.4 | 47.8 | 41.6 | 40.5 | 65.98 | 56.13 | 54.58 |
| DNNs | 82.3 | 71.6 | 68.7 | 68.3 | 57.4 | 55.8 | 52.5 | 46.2 | 44.4 | 40.4 | 38 | 36.6 | 60.88 | 53.3 | 51.38 |

Table: Phoneme error rate for various methods for "street" noise with different noise level

| | STREET | | | | | | | | | | | | Average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0DB | | | 5DB | | | 10DB | | | 15DB | | | | | |
| | Baseline | VTS | Proposed | Baseline | VTS | Proposed | Baseline | VTS | Proposed | Baseline | VTS | Proposed | Baseline | VTS | Proposed |
| TRI1 | 65.9 | 52.5 | 50.7 | 56.5 | 45.6 | 44.7 | 48.1 | 39.7 | 38.6 | 39.8 | 34.2 | 34.6 | 52.58 | 43 | 42.15 |
| TRI2 | 64 | 51.3 | 49.7 | 54.3 | 44 | 43.2 | 45.1 | 38.2 | 37.8 | 38 | 33.2 | 33.2 | 50.35 | 41.68 | 40.98 |
| TRI3 | 63.2 | 50.7 | 49.7 | 54.4 | 43.9 | 43.5 | 45.6 | 38.5 | 37.9 | 38.3 | 33.6 | 33.5 | 50.38 | 41.67 | 41.15 |
| DNNs | 59 | 48.4 | 47.4 | 48.1 | 41.4 | 40.8 | 41.5 | 36.2 | 35.4 | 34.7 | 32.2 | 31.9 | 45.83 | 39.55 | 38.87 |

# Conclusion

- We have proposed a new corruption function which includes effect of auditory masking.

# Conclusion

- We have proposed a new corruption function which includes effect of auditory masking.
- The proposed algorithms provide significant performance gain over the traditional VTS technique with little additional computational cost.

# Conclusion

- We have proposed a new corruption function which includes effect of auditory masking.
- The proposed algorithms provide significant performance gain over the traditional VTS technique with little additional computational cost.
- We are currently exploring methods to improve the MMSE estimation by introducing the clean model and compensated model variances into the estimation equation.

# THANK YOU