Stereo-Based Stochastic Noise Compensation Based on Trajectory GMMs

Heiga Zen Yoshihiko Nankaku Keiichi Tokuda (Nagoya Institute of Technology)

1. Introduction

GMM-based noise compensation

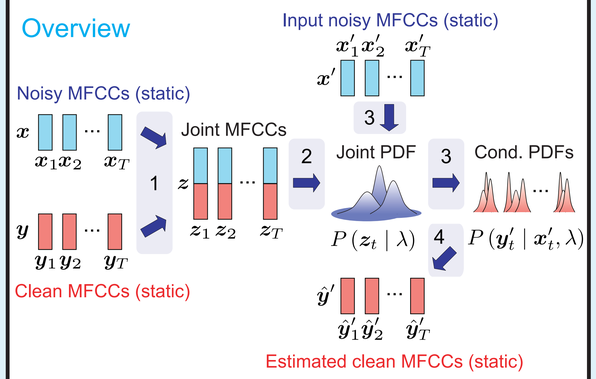
- Compensation process
  1. Prepare two-channel (noisy & clean) speech data
  2. Model joint PDF between clean & noisy MFCCs
  3. Estimate conditional PDFs of clean MFCCs
  4. Determine clean MFCCs
- Problem
  - Frame-by-frame mapping
    ⇒ Cannot use inter-frame information
  - Dynamic features or concatenating multiple frames
    ⇒ Improper in the sense of statistical modeling

Trajectory GMM-based noise compensation

- Trajectory HMM [Zen;'07]
  - Impose explicit relationships betw. static & dyn. feats.
    ⇒ HMM is reformulated as a trajectory model
  - Avoid limitations of HMMs w/o additional parameters
    * Frame-wise conditional independence assumption
    * Piece-wise constatnt statistics within an HMM state
- Noise compensation based on trajectory GMMs
  - Entire utterance-level mapping
    ⇒ Inter-frame correlation can be used
  - Using dyn. feat. constraints in both training & mapping
    ⇒ Proper statistical modeling

2. GMM-based noise compensation

Overview



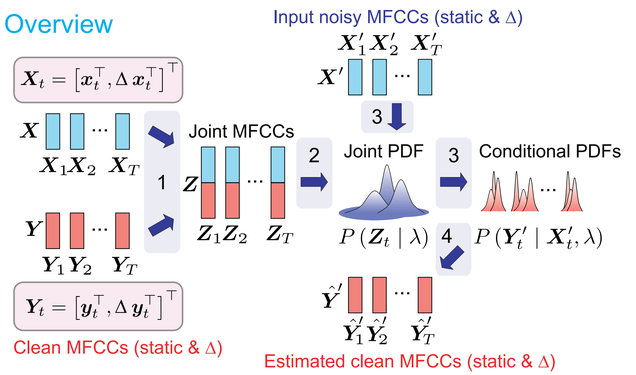
1. Make joint MFCC z_t from noisy & clean MFCCs x_t & y_t
2. Model frame-level joint PDF $P(z_t \mid \lambda)$ by a GMM
3. Convert joint PDF $P(z_t \mid \lambda)$ to cond. PDF $P(y'_t \mid x'_t, \lambda)$
4. Estimate clean MFCC $\hat{y}'_t$ from cond. PDF by MMSE

$$\hat{y}'_t = \sum_{i=1}^{N} \gamma_i \left[\mu_i^{(y)} + \Sigma_i^{(yx)} \Sigma_i^{(xx)^{-1}} \left(x'_t - \mu_i^{(x)} \right) \right]$$

Each frame is transformed independently
⇒ Inter-frame correlation are not used

3. GMM-based compensation with dyn. feats.

Overview



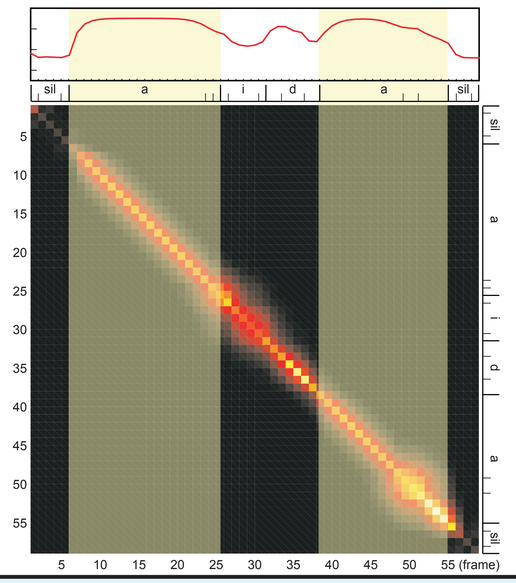
- Dynamic features (delta MFCCs) are also used
  ⇒ Inter-frame dependencies are also used
- Dynamic feature constraints are ignored
  ⇒ Dynamic-feature parts in mapped feats. no longer valid

4. Trajectory HMM

Trajectory HMM [Zen;'07]

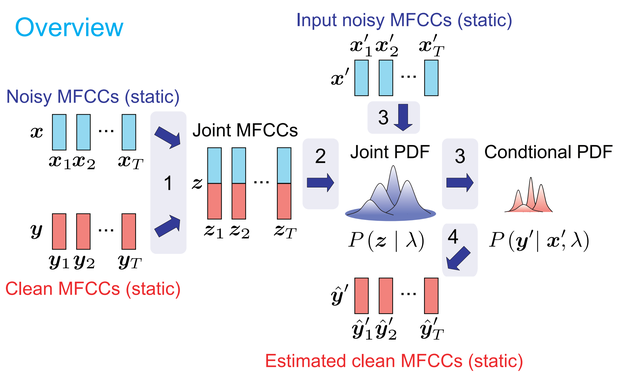
$$P(z \mid \lambda) = \sum_{\forall q} P(q \mid \lambda) P(z \mid q, \lambda) \qquad \mu_q = \left[\mu_{q_1}^\top, \ldots, \mu_{q_T}^\top \right]^\top$$
$$P(z \mid q, \lambda) = \mathcal{N}\left(z ;\, \bar{z}_q, P_q\right) \qquad \mu_i = \left[\mu_i^{(x)\top}, \mu_i^{(y)\top} \right]^\top$$
$$R_q \bar{z}_q = r_q \qquad \Sigma_q^{-1} = \mathrm{diag}\left[\Sigma_{q_1}^{-1}, \ldots, \Sigma_{q_T}^{-1} \right]$$
$$R_q = W^\top \Sigma_q^{-1} W = P_q^{-1} \qquad \Sigma_i^{-1} = \begin{bmatrix} \Omega_i^{(xx)} & \Omega_i^{(xy)} \\ \Omega_i^{(yx)} & \Omega_i^{(yy)} \end{bmatrix}$$
$$r_q = W^\top \Sigma_q^{-1} \mu_q$$

- $\bar{z}_q$ is determined as a *smooth trajectory*
  ⇒ Speech dynamics can be modeled explicitly
- Intra & inter-frame covariance matrix P_q is *full*
  ⇒ Capture both intra & inter-frame correlations



5. Trajectory GMM-based noise compensation

Overview



1. Make joint MFCCs z from clean & noisy MFCCs x & y

$$z = \left[z_1^\top, z_2^\top, \ldots, z_T^\top \right]^\top \qquad z_t = \left[x_t^\top, y_t^\top \right]^\top$$

2. Model utterance-level joint PDF $P(z \mid \lambda)$ by a trajectory GMM

$$P(z \mid \lambda) = \sum_{\forall q} P(q \mid \lambda) P(z \mid q, \lambda)$$
$$P(z \mid q, \lambda) = \mathcal{N}\left(z ;\, \bar{z}_q, P_q\right)$$

3. Convert joint PDF $P(z \mid \lambda)$ to conditional PDF $P(y' \mid x', \lambda)$

$$P(z \mid q, \lambda) = \mathcal{N}\left(\begin{bmatrix} x \\ y \end{bmatrix} ; \begin{bmatrix} \bar{x}_q \\ \bar{y}_q \end{bmatrix}, \begin{bmatrix} P_q^{(xx)} & P_q^{(xy)} \\ P_q^{(yx)} & P_q^{(yy)} \end{bmatrix} \right)$$
$$P(y' \mid x', \lambda) = \sum_{\forall q} \gamma_q \cdot P(y' \mid x', q, \lambda)$$
$$P(y' \mid x', q, \lambda) = \mathcal{N}\left(y' ;\, \tilde{y}_q, \tilde{P}_q^{(yy)}\right)$$
$$\tilde{y}_q = \bar{y}_q + P_q^{(yx)} C_q^{(xx)} \left(x' - \bar{x}_q \right)$$
$$\tilde{P}_q^{(yy)} = P_q^{(yy)} - P_q^{(yx)} C_q^{(xx)} P_q^{(xy)}$$

4. Estimate clean MFCCs $\hat{y}'$ from conditional PDF by MMSE

$$\hat{y}'_{\mathrm{MMSE}} = \sum_{\forall q} \gamma_q \left[\bar{y}_q + P_q^{(yx)} C_q^{(xx)} \left(x' - \bar{x}_q \right) \right]$$

- Entire utterance-level transform
  ⇒ Inter-frame correlation are used in mapping
- Dyn. feat. constraints are used in both training & mapping
  ⇒ Make training & mapping consistent

6. Speech recognition experiment

Conditions

HMMs for recognition

Database	AURORA-2 database
Training data	Clean-condition training data
Feature vec.	12 MFCC & log energy by ETSI front-end 2.0, Δ & ΔΔ
Topology	16-state 3-mix left-to-right HMM for each digit

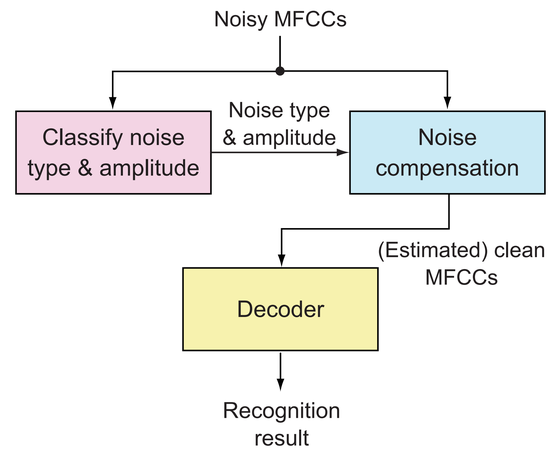
GMMs/trajectory GMMs for noise compensation

Database	AURORA-2 database
Training data	Multi-condition training data
Feature vec.	12 MFCC & log energy by ETSI front-end 2.0 (Δ & ΔΔ)
Topology	(Trajectory) GMM: 256-mix for each noise condition

GMMs for noise classification

Database	AURORA-2 database
Training data	Multi-condition training data, first 10 frames only
Feature vec.	12 MFCC & log energy by ETSI front-end 2.0, Δ & ΔΔ
Topology	GMM: 4-mix for each noise condition

Recognition process



Experimental results (average word acc. (%))

Mapping	Set A	Set B	Set C	Average
w/o compensation	61.34	55.75	66.14	60.06
GMM (static)	82.49	74.32	69.30	76.58
GMM (static & Δ)	88.03	**82.80**	77.92	83.92
Trajectory GMM	**89.38**	81.54	**80.87**	**84.54**

Set A: seen noise, Set B: unseen noise, Set C: channel mismatch

- Proposed technique worked better for seen noise but worse for unseen noise
  ⇒ Overfitting due to dynamic feature constraints?