Stereo-Based Stochastic Mapping for Robust speech Recognition

Author : Mohamed Afify, Xiaodong Cui
Yuqing Gao

Professor:陳嘉平
Reporter:葉佳璋

Outline

- Introduction
- Algorithm Formulation
- Comparison Between The Method And Other Similar Technique
- Experiment

Introduction

- We present a stochastic mapping technique for robust speech recognition that use stereo data.

- The idea is based constructing a Gaussian mixture model for the joint distribution of the clean and noisy features .

stereo-based stochastic mapping

- The proposed transformation is built using stereo data, i.e., data that consists of simultaneous recordings of both the clean and noisy speech.
- We will refer to this mapping as stereo-based stochastic mapping(SSM).

Algorithm Formulation

- Assume we have a set of stereo data $\{(x_i, y_i)\}$
  - ➢x: clean feature representation of speech
  - ➢y: noisy feature representation of speech
  - ➢N: be the number of these feature vectors $(1 \le i \le N)$
  - ➢M: feature vector dimension
- Define $z \equiv (x, y)$ concatenation of the two channels.

Joint Probability

- The first step in constructing the mapping is training the joint probability *p(z).*

$$p(z) = \sum_{k=1}^{K} c_k \mathcal{N}(z; \mu_{z,k}, \Sigma_{zz,k}) \quad (1)$$

  - K : number of mixture components
  - c_k : weights of components
  - $\mu_{z,k}$: mean of components
  - $\Sigma_{zz,k}$: covariance of components

Joint Probability(cont.)

- In the most general case
  - L_n : noisy vectors
  - L_c : clean vectors
  - $M(L_n + L_c)$: the size of z and $\mu_{z,k}$
  - $M(L_n + L_c) \times M(L_n + L_c)$:the size of $\sum_{zz,k}$
- The mean and covariance can be partition as

$$\mu_{z,k} = \begin{pmatrix} \mu_{x,k} \\ \mu_{y,k} \end{pmatrix} \quad (2) \qquad \sum_{zz,k} = \begin{pmatrix} \sum_{xx,k} & \sum_{xy,k} \\ \sum_{yx,k} & \sum_{yy,k} \end{pmatrix} \quad (3)$$

MAP-Based Estimation

- The clean feature x give the noise observation y be formulated as

$$
\begin{aligned}
\hat{x} &= \arg\max_x p(x \mid y) \\
&= \arg\max_x \sum_k p(x, k \mid y) \\
&\equiv \arg\max_x \log \sum_k p(x, k \mid y) \quad (4)
\end{aligned}
$$

- Define the log likelihood as

$$
L(x) \equiv \log \sum_k p(x, k \mid y) \quad (5)
$$

MAP-Based Estimation(cont.)

- Auxiliary function

$$Q(x,\overline{x}) \equiv \sum_k p(k \mid \overline{x}, y) \log p(x, k \mid y) \quad (6)$$

- auxiliary objective function proceed at each iteration as follows:

$$\hat{x} = \arg\max_x \sum_k p(k \mid \overline{x}, y) \log\left(p(k \mid y) p(x \mid k, y) \right)$$

$$= \arg\max_x \sum_k p(k \mid \overline{x}, y) \left[\log p(k \mid y) + \log p(x \mid k, y) \right]$$

$$\equiv \arg\max_x \sum_k p(k \mid \overline{x}, y) \log p(x \mid k, y)$$

$$\equiv \arg\max_x \frac{-1}{2} \sum_k p(k \mid \overline{x}, y) \left[\log \left| \Sigma_{x|y,k} \right| + (x - \mu_x)^T \Sigma_{x|y,k}^{-1} (x - \mu_x) \right] \quad (7)$$

MAP-Based Estimation(cont.)

- By differentiating with respect x, setting the derivative to zero

$$\sum_k p(k \mid \bar{x}, y) \Sigma_{x|y,k}^{-1} \hat{x} = \sum_k p(k \mid \bar{x}, y) \Sigma_{x|y,k}^{-1} \mu_{x|y,k} \quad (8)$$

- And the condition statistic are known to be

$$\mu_{x|y,k} = \mu_{x,k} + \Sigma_{xy,k} \Sigma_{yy,k}^{-1} (y - \mu_{y,k}) \quad (9)$$

$$\Sigma_{x|y,k} = \Sigma_{xx,k} - \Sigma_{xy,k} \Sigma_{yy,k}^{-1} \Sigma_{yx,k} \quad (10)$$

MAP-Based Estimation(cont.)

- The mapping (8)-(10) can be rewrite as a mixture of linear transformations weighted by component posteriors as follows

$$\hat{x} = \sum_k p(k \mid \bar{x}, y)(A_k y + b_k) \quad (12)$$

$$where\ \mathrm{A_k} = CD_k,\ \mathrm{b_k} = Ce_k$$

$$C = \left(\sum_k p(k \mid \bar{x}, y) \Sigma_{x|y,k}^{-1} \right)^{-1} \quad (13)$$

$$D_k = \Sigma_{x|y,k}^{-1} \Sigma_{yy,k}^{-1} \Sigma_{xy,k} \quad (14)$$

$$e_k = \Sigma_{x|y,k}^{-1} \left(\mu_{x,k} - \Sigma_{yy,k}^{-1} \Sigma_{xy,k} \mu_{y,k} \right) \quad (15)$$

MMSE-Based Estimation

- The clean feature x give the noise speech feature y

$$\hat{x} = E[x \mid y] \quad (16)$$

- Considering the GMM structure of the joint distribution, can be further decomposed as

$$\hat{x} = \int_x p\left(x \mid y\right) x dx = \sum_k \int_x p\left(x, k \mid y\right) x dx$$

$$= \sum_k p\left(k \mid y\right) \int_x p\left(x \mid k, y\right) x dx$$

$$= \sum_k p\left(k \mid y\right) E[x \mid k, y] \quad (17)$$

MMSE-Based Estimation(cont.)

- In (17), the posterior probability term p(k|y) can be computed as

$$p(k \mid y) = \frac{p(k, y)}{p(y)} = \frac{p(y \mid k)\, p(k)}{\sum_k p(y \mid k) p(k)} \quad (18)$$

- And the expectation term E[x|k,y] is given in (9).

MMSE-Based Estimation(cont.)

- Also the MMSE predictor can be written as weighted sum of linear transformations as follows:

$$\hat{x} = \sum_k p\left(k \mid y\right)(F_k y + g_k) \quad (19)$$

where

$$F_k = \Sigma_{xy,k} \Sigma_{yy,k}^{-1} \quad (20)$$

$$\mathrm{g_k} = \mu_{x,k} - \Sigma_{xy,k} \Sigma_{yy,k}^{-1} \mu_{y,k} (21)$$

Relationships Between MAP and MMSE Estimators

- To highlight the iterative nature of the MAP estimator

$$\hat{x}^l = \sum_k p(k \mid \overline{x}^{l-1}, y)(A_k y + b_k) \quad (22)$$

- If we compare one iteration of (22) to (19)
  - MAP uses a posterior $p(k \mid \overline{x}^{l-1}, y)$ calculated from the joint probability distribution.
  - MMSE employs a posteriors $p(k \mid y)$ based on the marginal probability distribution.

Relationships Between MAP and MMSE Estimators(cont.)

- If we compare to coefficients of the transformations in (13)-(15) and (20)-(21).
- We can see MAP has extra term

$$\left(\sum_k p(k \mid \bar{x}^{l-1}, y) \Sigma_{x|y,k}^{-1} \right)^{-1} \quad (23)$$

Relationships Between MAP and MMSE Estimators(cont.)

- If we assume the conditional covariance matrix in (23) is constant across k,

$$\left(\sum_k p(k \mid \overline{x}^{l-1}, y)\,\Sigma_{x|y,k}^{-1}\right)^{-1}$$

$$= \left(\Sigma_{x|y}^{-1} \sum_k p(k \mid \overline{x}^{l-1}, y)\right)^{-1} = \left(\Sigma_{x|y}^{-1} \cdot 1\right)^{-1} = \Sigma_{x|y} \quad (24)$$

- (25) and (26) are the same to (20) and (21)

$$\mathrm{A_k} = \Sigma_{x|y}\,\Sigma_{x|y}^{-1}\left(\Sigma_{xy,k}\,\Sigma_{yy,k}^{-1}\right)$$

$$= \Sigma_{xy,k}\,\Sigma_{yy,k}^{-1} \quad (25)$$

$$\mathrm{b_k} = \Sigma_{x|y}\,\Sigma_{x|y}^{-1}\left(\mu_{x,k} - \Sigma_{xy,k}\,\Sigma_{yy,k}^{-1}\,\mu_{y,k}\right)$$

$$= \mu_{x,k} - \Sigma_{xy,k}\,\Sigma_{yy,k}^{-1}\,\mu_{y,k} \quad (26)$$

Comparison The Proposed Method And other similar technique.

- The proposed method is effectively a mixture of linear transformations weighted by component posteriors.

- This is similar to several recently proposed algorithms.
  - SPLICE
  - CMLLR

SPLICE

- SPLICE is a recently proposed noise compensation algorithm that uses stereo data.

$$\hat{x} = \sum_k p\left(k \mid y\right)\left(y + r_k\right) \quad (27)$$

where

$$r_k = \frac{\sum_n p\left(k \mid y_n\right)\left(x_n - y_n\right)}{\sum_n p\left(k \mid y_n\right)} \quad (28)$$

and n run over the data.

Compare SPLICE to MMSE

- Compare to MMSE-based SSM we can observe
  - First, SPLICE builds a GMM on noisy features while in this paper a GMM is built on the joint clean and noisy features (1).
  - Second, SPLICE is a special case of SSM if the clean and noisy feature are perfectly correlated. ($\sum_{xy,k} = \sum_{yy,k}$, and $\mathrm{p}(\mathrm{k} \mid \mathrm{x}_{\mathrm{n}}) = \mathrm{p}(\mathrm{k} \mid \mathrm{y}_{\mathrm{n}})$)

$$r_k = \frac{\sum_n p(k \mid y_n)(x_n - y_n)}{\sum_n p(k \mid y_n)}$$

$$= \frac{\sum_n p(k \mid y_n) x_n - \sum_n p(k \mid y_n) y_n}{\sum_n p(k \mid y_n)} = \frac{\sum_n p(k \mid y_n) x_n}{\sum_n p(k \mid y_n)} - \frac{\sum_n p(k \mid y_n) y_n}{\sum_n p(k \mid y_n)}$$

$$= \frac{\sum_n p(k \mid x_n) x_n}{\sum_n p(k \mid y_n)} - \frac{\sum_n p(k \mid y_n) y_n}{\sum_n p(k \mid y_n)} = \mu_{x,k} - \mu_{y,k}$$

CMLLR

- There are several recently proposed technique use a mixture of CMLLR transforms.
- These can be written as

$$\hat{x} = \sum_{k} p\left(k \mid y\right)(U_k y + v_k) \quad (27)$$

Where

p(k|y) is calculated using an auxiliary Gaussian mixture model that is train on noisy observation.

U_k and v_k are the elements of CMLLR that do not require stereo data for their estimation.

SSM and CMLLR-Based Method

- The major difference between SSM and the previous methods lies in the used GMM (again noisy and versus joint).

- SSM is similar in principle to training-based techniques and can be combined with adaptation methods.

Experiments

- Large-vocabulary spontaneous English speech recognition task.
  - The original (clean) training data 150 h of speech. This data is used to build the clean acoustic model.
  - MST(multi style training) Model is also trained from the MST data.
  - The noisy data are generated by adding humvee, tank, and babble noise.
  - The experiment are carried out on two test set.
  - Set A: utterance recorded in the clean condition, and are corupted artificially noise to produce 15-db and 10-db noisy test data.
  - Set B: utterance recorded in a real world, and the SNRs are measured around 5 db to 10 db.

Experimental Results

WORD ERROR RATE RESULTS (IN %) OF THE COMPENSATION SCHEMES AGAINST CLEAN ACOUSTIC MODEL

	Set A			Set B
	clean	15 dB	10 dB	5-8 dB
clean model	4.84	18.40	33.66	47.72
clean model + CMLLR	3.23	14.30	27.89	43.28
SSM_MAP1	4.87	18.05	33.32	48.24
SSM_MAP1 + CMLLR	3.23	14.41	28.79	43.63
SSM_MAP3	4.87	18.03	33.36	46.04
SSM_MAP3 + CMLLR	3.23	14.43	28.36	41.68
SSM_MMSE	4.84	13.39	25.52	28.43
SSM_MMSE + CMLLR	3.26	13.23	25.12	28.25
SSM_MMSE_MAP3	4.84	13.12	25.26	28.14
SSM_MMSE_MAP3 + CMLLR	3.23	12.17	23.76	27.07

MST

WORD ERROR RATE RESULTS (IN %) OF THE COMPENSATION SCHEMES AGAINST MST ACOUSTIC MODEL

	Set A			Set B
	clean	15 dB	10 dB	5-8 dB
MST model	7.67	11.06	18.90	46.74
MST model + CMLLR	3.87	7.69	14.13	25.87
SSM_MAP1	4.57	9.75	18.46	43.59
SSM_MAP1 + CMLLR	2.74	6.96	14.07	23.83
SSM_MAP3	4.77	9.32	17.59	40.58
SSM_MAP3 + CMLLR	2.76	6.79	13.78	22.85
SSM_MMSE	4.15	10.41	20.39	31.57
SSM_MMSE + CMLLR	2.76	8.50	17.66	18.31
SSM_MMSE_MAP3	3.96	9.66	19.20	26.49
SSM_MMSE_MAP3 + CMLLR	2.74	7.70	16.23	15.95

$$E(x\,|\,y)=\int p(x\,|\,y)xdx=\int \frac{p(x,y)}{p(y)}xdx=\int \frac{c^{'}e^{-\frac{1}{2}(z-\mu_z)^T\sum_{zz}(z-\mu_z)}}{ce^{-\frac{1}{2}(y-\mu_y)^T\sum_{yy}(y-\mu_y)}}xdx=\int Ce^{-\frac{1}{2}(x-\mu_{x|y})^T\sum_{x|y}(x-\mu_{x|y})}xdx$$

$$(x-\mu_x, y-\mu_y)\begin{pmatrix}\sum_{xx} & \sum_{xy}\\ \sum_{yx} & \sum_{yy}\end{pmatrix}^{-1}\begin{pmatrix}x-\mu_x\\ y-\mu_y\end{pmatrix}-(y-\mu_y)\sum_{yy}{}^{-1}(y-\mu_y)$$

$$=(a_i,b_i)\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_2\sigma_1 & \sigma_2^2\end{pmatrix}^{-1}\begin{pmatrix}a_i\\ b_i\end{pmatrix}-(b_i)(\sigma_2^2)^{-1}(b_i)$$

$$=\frac{1}{\sigma_1^2\sigma_2^2-\rho\sigma_1\sigma_2\rho\sigma_2\sigma_1}(a_i,b_i)\begin{pmatrix}\sigma_2^2 & -\rho\sigma_1\sigma_2\\ -\rho\sigma_1\sigma_2 & \sigma_1^2\end{pmatrix}\begin{pmatrix}a_i\\ b_i\end{pmatrix}-(b_i)(\sigma_2^2)^{-1}(b_i)$$

$$\frac{1}{\sigma_1^2\sigma_2^2-\rho\sigma_1\sigma_2\rho\sigma_2\sigma_1}(a_i,b_i)\begin{pmatrix}\sigma_2^2 & -\rho\sigma_1\sigma_2\\ -\rho\sigma_1\sigma_2 & \sigma_1^2\end{pmatrix}\begin{pmatrix}a_i\\ b_i\end{pmatrix}-(b_i)(\sigma_2^2)^{-1}(b_i)$$

$$=\frac{1}{\sigma_1^2\sigma_2^2-\rho^2\sigma_1^2\sigma_2^2}\left[\left(a_i\sigma_2^2-b_i\rho\sigma_1\sigma_2,-a_i\rho\sigma_1\sigma_2+b_i\sigma_1^2\right)\begin{pmatrix}a_i\\ b_i\end{pmatrix}-\left(\sigma_1^2\sigma_2^2-\rho^2\sigma_1^2\sigma_2^2\right)(b_i)^2(\sigma_2^2)^{-1}\right]$$

$$=\frac{\left(\sigma_2^2\right)^{-1}}{\sigma_1^2-\rho^2\sigma_1^2}\left[\left(a_i{}^2\sigma_2^2-a_ib_i\rho\sigma_1\sigma_2-a_ib_i\rho\sigma_1\sigma_2+b_i{}^2\sigma_{1\,i}^2\right)-\left((b_i)^2(\sigma_2^2)^{-1}\sigma_1^2\sigma_2^2-(b_i)^2(\sigma_2^2)^{-1}\rho^2\sigma_1^2\sigma_2^2\right)\right]$$

$$=\frac{1}{\sigma_1^2-\rho^2\sigma_1^2}\left[\left(\left(\sigma_2^2\right)^{-1}a_i{}^2\sigma_2^2-\left(\sigma_2^2\right)^{-1}a_ib_i\rho\sigma_1\sigma_2-\left(\sigma_2^2\right)^{-1}a_ib_i\rho\sigma_1\sigma_2+\left(\sigma_2^2\right)^{-1}b_i{}^2\sigma_1^2\right)-\left(\left(\sigma_2^2\right)^{-1}(b_i)^2\sigma_1^2-\left(\sigma_2^2\right)^{-1}(b_i)^2\rho^2\sigma_1^2\right)\right]$$

$$=\frac{1}{\sigma_1^2-\rho^2\sigma_1^2}\left[\left(a_i{}^2-2a_ib_i\rho\sigma_1\sigma_2^{-1}+\left(b_i\sigma_1\sigma_2^{-1}\right)^2\right)-\left(b_i\sigma_1\sigma_2^{-1}\right)^2(1-\rho^2)\right]$$

$$=\frac{1}{\sigma_1^2-\rho^2\sigma_1^2}\left[\left(a_i{}^2-2a_ib_i\rho\sigma_1\sigma_2^{-1}+\left(b_i\rho\sigma_1\sigma_2^{-1}\right)^2\right)\right]$$

$$=\left(a_i-b_i\rho\sigma_1\sigma_2^{-1}\right)\frac{1}{\sigma_1^2-\rho^2\sigma_1^2}\left(a_i-b_i\rho\sigma_1\sigma_2^{-1}\right)$$

$$\left(a_i - b_i \rho \sigma_1 \sigma_2^{-1}\right) \frac{1}{\sigma_1^2 - \rho^2 \sigma_1^2} \left(a_i - b_i \rho \sigma_1 \sigma_2^{-1}\right)$$

$$= \left(x - \mu_x - \Sigma_{xy} \Sigma_{yy}{}^{-1} \left(y - \mu_y\right)\right)^T \left(\Sigma_{xx,k} - \Sigma_{xy,k} \Sigma_{yy,k}^{-1} \Sigma_{yx,k}\right)^{-1} \left(x - \mu_x - \Sigma_{xy} \Sigma_{yy}{}^{-1} \left(y - \mu_y\right)\right)$$

$$\mu_{x|y,k} = \mu_{x,k} + \Sigma_{xy,k} \Sigma_{yy,k}^{-1} (y - \mu_{y,k}) \quad (9)$$

$$\Sigma_{x|y,k} = \Sigma_{xx,k} - \Sigma_{xy,k} \Sigma_{yy,k}^{-1} \Sigma_{yx,k} \quad (10)$$

MAP-Based Estimation(cont.)

- Special case arise when x is a scalar.
  - Use ith noisy coefficient to predict the clean coefficient.
  - Use a time window around the ith noisy coefficient to predict.

- The solution in (8) will reduces to the following for every vector dimension:

$$\hat{x} = \frac{\sum_k p(k \mid \bar{x}, y)\mu_{x|y,k} / \sigma^2_{x|y,k}}{\sum_k p(k \mid \bar{x}, y)\Sigma^{-1}_{x|y,k} / \sigma^2_{x|y,k}} \quad (11)$$

- Where $\sigma^2_{x|y,k}$ is used instead of $\Sigma_{x|y,k}$ to indicate that it is a scalar.