Efficient VQ-based MMSE Estimation for Robust Speech Recognition

J.A. González, A.M. Peinado, A.M. Gómez, J.L. Carmona, and J.A. Morales-Cordovilla

Dpt. de Teoría de la Señal, Telemática y Comunicaciones

Abstract

This paper presents a feature compensation technique based on the minimum mean square error (MMSE) estimation for robust speech recognition. Similarly to other MMSE compensation methods based on stereo data, our approach models the differences between clean and noisy feature spaces, and the resulting MMSE estimate of the clean feature vector is obtained as a piece-wise linear transformation of the noisy one. Unlike other well-known MMSE techniques such as SPLICE or MEMLIN, which model the feature spaces with Gaussian mixture models (GMMs), in our proposal each feature space is characterized by a set of cells obtained by means of vector quantization (VQ). This VQ-based approach allows a very efficient implementation of the MMSE estimator. Also, the inherent degradation of any VQ process is overcome by a strategy based on considering different subregions inside each cell and a subregion-based mean and variance compensation. The experimental results show that, along with a very efficient MMSE estimator, our technique achieves even better recognition accuracies than SPLICE and MEMLIN.

Introduction

- **Noise robustness** is a crucial component of the speech-enabled applications for mobile devices.
- **Methods for robust speech recognition:**
    - Feature-domain techniques.
    - Model-domain techniques.
- **Feature-domain techniques** modify or enhance the input test data to be closer to the clean training data.
- **Our proposal:** feature compensation technique based on MMSE estimation and stereo-data.
- **Stereo database:** contains clean and noisy feature vectors to learn the statistical relationship between the clean and noisy feature spaces.

MMSE estimation

The MMSE estimate of the clean feature vector $\boldsymbol{x}$ given the distorted input $\boldsymbol{y}$ can be computed as,

$$\hat{\boldsymbol{x}} = E[\boldsymbol{x}|\boldsymbol{y}] = \int_{\boldsymbol{x}} \boldsymbol{x} \cdot p(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x}$$

How can we model $p(\boldsymbol{x}|\boldsymbol{y})$?

We assume that the clean and distorted feature spaces can be represented by means of ***pdf mixtures***. Thus,

$$p(\boldsymbol{x}|\boldsymbol{y}) = \sum_{k_x} \sum_{k_y} p\left(\boldsymbol{x}|k_x, k_y, \boldsymbol{y}\right) p\left(k_x|k_y, \boldsymbol{y}\right) p\left(k_y|\boldsymbol{y}\right)$$

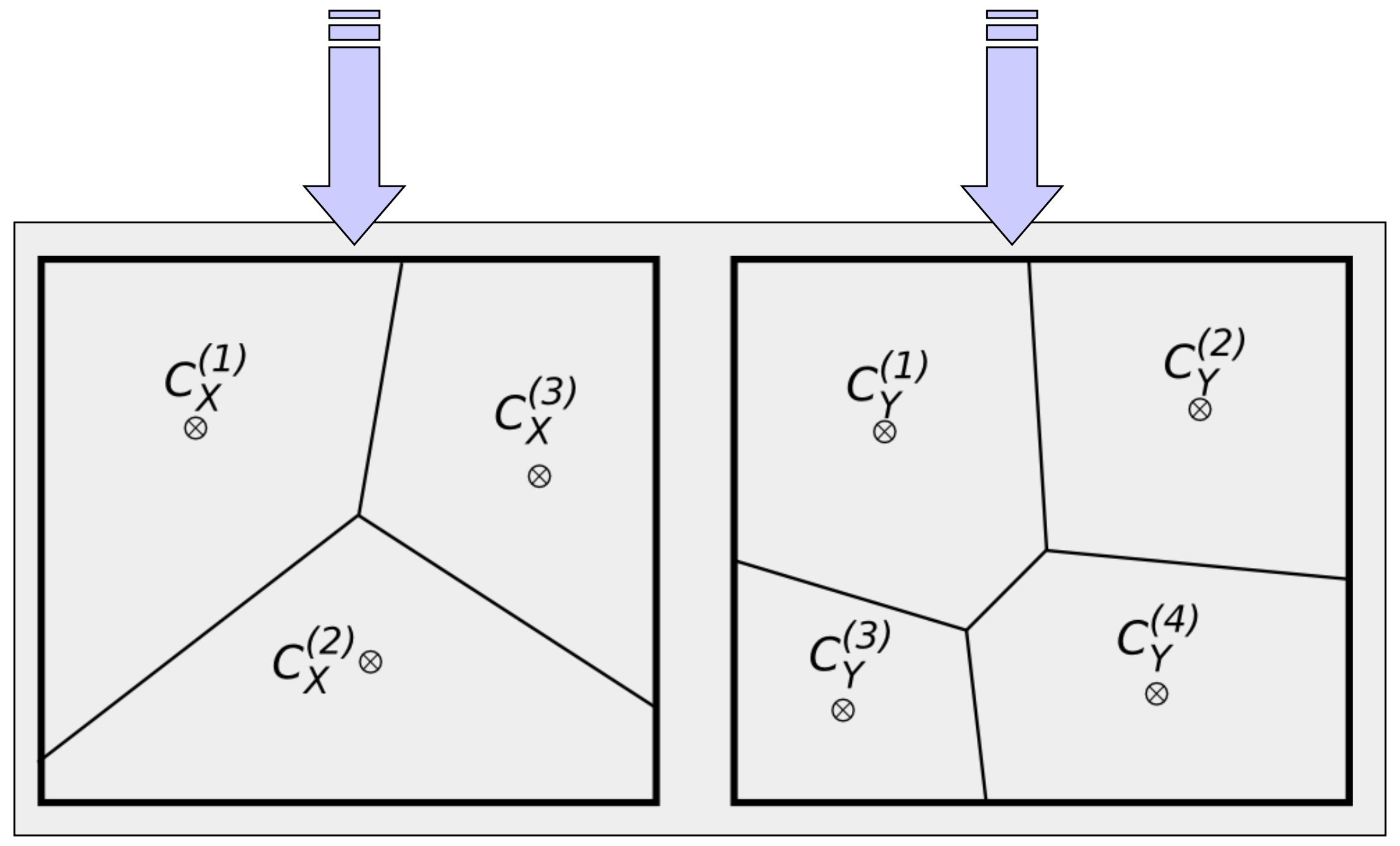
The ***MMSE estimate*** takes now the following form,

$$\hat{\boldsymbol{x}} = \sum_{k_x} \sum_{k_y} E\left[\boldsymbol{x}|k_x, k_y, \boldsymbol{y}\right] P\left(k_x|k_y, \boldsymbol{y}\right) P\left(k_y|\boldsymbol{y}\right)$$

Our proposal: VQ-based MMSE estimation

We model the clean and distorted feature spaces by means VQ codebooks. These codebooks partition each feature space into a set of cells:

$$C_X^{(i)} (i = 1, \ldots, M) \qquad C_Y^{(j)} (j = 1, \ldots, N)$$



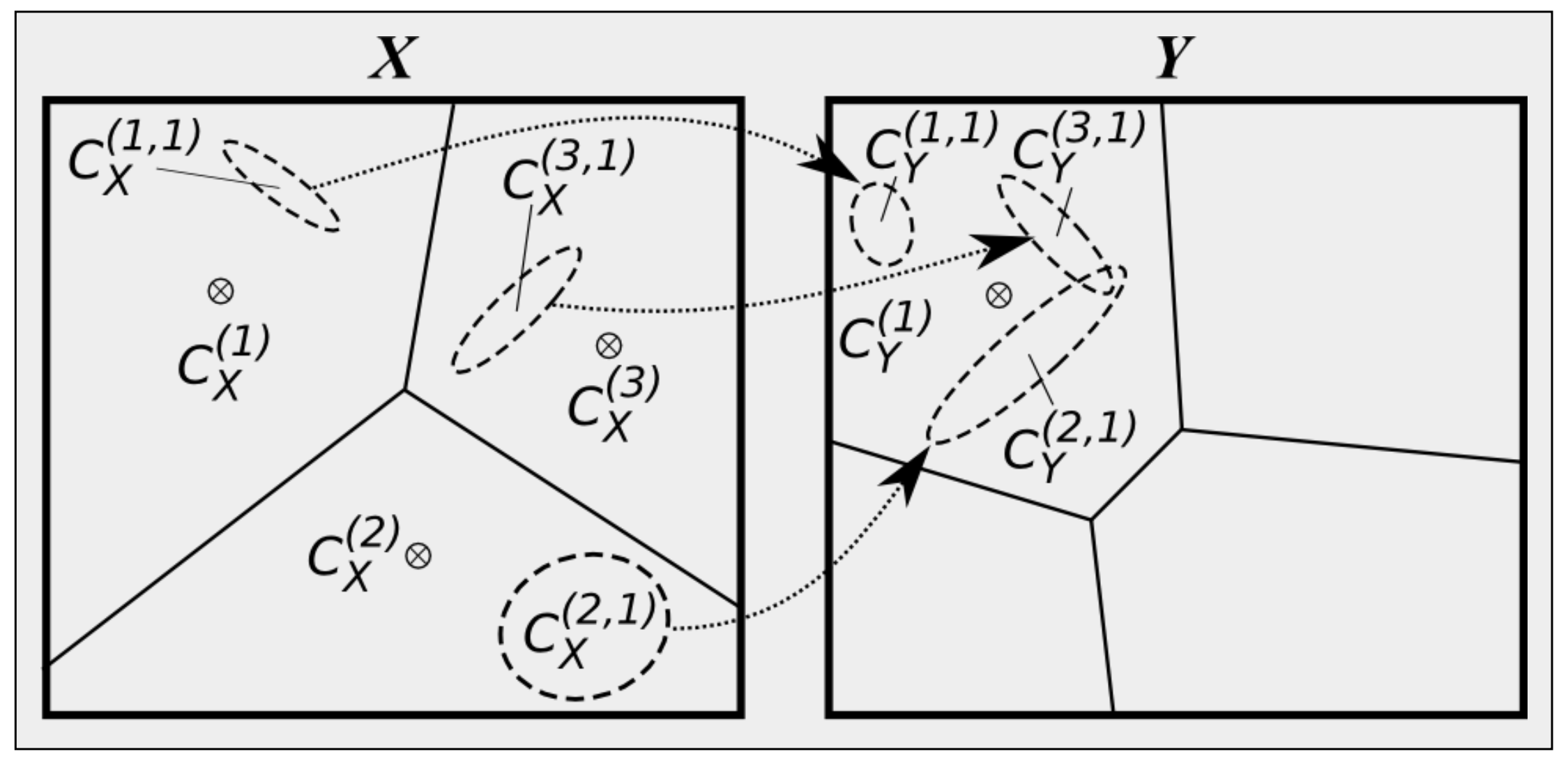
The ***proposed MMSE estimation*** is,

$$\hat{\boldsymbol{x}} = \sum_{i=1}^{M} E\left[\boldsymbol{x} \middle| C_X^{(i)}, C_Y^{*}, \boldsymbol{y}\right] P\left(C_X^{(i)} \middle| C_Y^{*}\right)$$

where $E[\boldsymbol{x}|C_X^{(i)}, C_Y^{*}, \boldsymbol{y}]$ defines the mapping between the clean cell $C_X^{(i)}$ and the noisy one C_Y^{*} due to environmental noise.

Improving the hard VQ modeling

We introduce the concept of ***subregions*** in a VQ cell:



Thus, the transformation between Gaussian distributed noisy and clean subregions is given by,

$$E\left[\boldsymbol{x} \middle| C_X^{(i)}, C_Y^{(j)}, \boldsymbol{y}\right] = \boldsymbol{\mu}_X^{(i,j)} + \left(\boldsymbol{\Sigma}_X^{(i,j)}\right)^{1/2} \left(\boldsymbol{\Sigma}_Y^{(i,j)}\right)^{-1/2} \left(\boldsymbol{y} - \boldsymbol{\mu}_Y^{(i,j)}\right)$$

Other MMSE estimators

SPLICE:

$$\hat{\boldsymbol{x}} = \sum_{k_y} \left(\boldsymbol{y} - \boldsymbol{r}_{k_y}\right) P(k_y|\boldsymbol{y})$$

MEMLIN:

$$\hat{\boldsymbol{x}} = \sum_{k_y} \sum_{k_x} \left(\boldsymbol{y} - \boldsymbol{r}_{k_x,k_y}\right) P(k_x|k_y) P(k_y|\boldsymbol{y})$$

Performance Evaluation

- **Word Accuracy results** (WAcc %).
- **Task:** recognition of connected English digits under artificially added acoustic noise (based on the **Aurora-2** database).
- **Environmental conditions:** 55 environments (9 noises at 6 SNRs plus the clean condition).
- **Speech feature extractor**: ETSI Front-end.
- VQ codebooks with **256 cells** are employed.
- Diagonal covariance GMMs with **256 Gaussians** are used by MEMLIN and SPLICE.

Results

Oracle results: the test environment is known.

	20 dB	15 dB	10 dB	5 dB	0 dB	Avg.
Baseline	90.79	75.53	50.70	25.86	11.27	**50.83**
SPLICE	98.09	95.87	88.88	70.62	39.04	**78.50**
MEMLIN	98.36	97.01	92.43	78.26	47.03	**82.62**
iVQ-MMSE	98.23	96.79	91.60	76.82	46.60	**82.01**
dVQ-MMSE	98.33	97.06	92.43	78.70	48.88	**83.08**
fVQ-MMSE	98.37	97.15	92.88	79.61	50.04	**83.61**

Soft-compensation results: the final estimate is computed as a weighted combination of the compensations for the learned environments.

$$\hat{\boldsymbol{x}} = \sum_{e} \hat{\boldsymbol{x}}_e P(e|\boldsymbol{y})$$

	Wacc (%)
SPLICE	72.99
MEMLIN	**77.21**
iVQ-MMSE	77.29
dVQ-MMSE	79.04
fVQ-MMSE	**79.54**

Conclusions

In this paper we have proposed a novel feature compensation technique based on MMSE estimation. Our proposal is shown to be a piece-wise linear function between the noisy feature space an the clean one, both modeled by means of VQ codebooks. The VQ modeling allows an efficient implementation of the MMSE estimator. In addition, a novel subregion-based approach is proposed in order to reduce the degradations introduced by the VQ quantization. The results show that the proposed technique can achieve better recognition accuracy than other well-known MMSE-based estimators (SPLICE and MEMLIN). Furthermore, these results show the importance of modeling both feature spaces in order to obtain a more accurate probability model for the MMSE estimation. In addition, the compensation of feature vectors taking into account the transformations in mean and covariance introduced by the noise leads to further improvements.