YET ANOTHER GAUSSIAN MIXTURE MODEL-BASED FEATURE COMPENSATION METHOD FOR ROBUST NOISY-DIGIT RECOGNITION

Chia-Ping Chen and Bing-Feng Yeh

National Sun Yat-sen University
Department of Computer Science and Engineering
70, Lien-Hai Road, Kaohsiung, Taiwan 804

ABSTRACT

We propose yet another Gaussian mixture model (YGMM) for robust speech recognition in noisy environments. The main difference between the proposed method and previously proposed GMM-based methods is that we estimate the noise features instead of the clean-speech features. In the implemented system, a condition classifier, incidentally based on GMM, is used to decide the noise type and level, and the corresponding GMM is employed to compensate for the noise-corrupted features. The proposed method and the implemented system are evaluated with the well-documented Aurora 2.0 noisy digit corpus. The results are promising. Specifically, it achieves a relative improvement in word error rate of 52.4% over the standard baseline, and 24.9% over a better baseline based on a traditional GMM-based feature compensation method.

Index Terms— Gaussian mixture model, Aurora 2.0, noise-robust speech recognition

1. INTRODUCTION

The immense popularity of *Google voice search* and *iPhone Siri*, among other internet-based services in recent years, makes strong case for the maturity of automatic speech recognition (ASR) technology. However, the issue of noise-robust speech recognition remains challenging for various applications. That is, the mismatch between the train data and the test data often leads to severe performance degradation, especially in noisy test environments. This is a critical problem to solve as noises are abundant in everyday lives.

Traditional methods have been proposed to achieve noise robustness. In speech enhancement, common approaches include spectral subtraction [1] and Wiener filtering [2]. In robust feature extraction, common approaches are cepstral mean subtraction [3], cepstral variance normalization [4], and histogram equalization [5]. More recently, exemplar-based methods, such as sparse representation, audio implanting, and compressive sensing, have been studied [6, 7, 8] with good performance. Features based on Teager-Kaiser energy estimator have also been investigated [9].

In this paper, we propose a noise-robust feature compensation method based on Gaussian mixture models (GMMs). Gaussian mixture models have been widely used on spoken language technology such as speaker recognition [10, 11, 12, 13] and voice conversion [14, 15, 16, 17]. One class of of GMMs used to model parallel data, also known as stereo data, have been successful in noise-robust recognition [18, 19, 20]. Recently, an implementation takes into account the dependence between adjacent frames [21], thus enforcing the inherent continuity constraints of speech features.

This paper is organized as follows. Traditional and the proposed methods using Gaussian mixture models for parallel data are introduced in Section 2. The evaluation scheme, results, and comments are described in Section 3. Concluding remarks and feature works are summarized in Section 4. Additionally, the relation of this work to prior work is provided in Section 5.

2. GAUSSIAN MIXTURE MODELS

2.1. Notation

Let $\mathbf{x}$ (column vector) denote a noisy speech feature vector of dimension D, $\mathbf{y}$ denote the corresponding clean speech feature vector of dimension D, and

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \tag{1}$$

be the concatenated feature vector of dimension $2D$. A Gaussian mixture model with K mixture components on $\mathbf{z}$ is denoted by $\mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_{\mathbf{z},k}, \boldsymbol{\Sigma}_{\mathbf{z},k}),\ k = 1, \ldots, K$, where

$$\boldsymbol{\mu}_{\mathbf{z},k} = \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{x},k} \\ \boldsymbol{\mu}_{\mathbf{y},k} \end{bmatrix}, \quad \boldsymbol{\Sigma}_{\mathbf{z},k} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{xx},k} & \boldsymbol{\Sigma}_{\mathbf{xy},k} \\ \boldsymbol{\Sigma}_{\mathbf{yx},k} & \boldsymbol{\Sigma}_{\mathbf{yy},k} \end{bmatrix}. \tag{2}$$

It can be shown (e.g. [22] which contains a practical introduction of Gaussian models) that the marginal and the conditional probability density functions of a joint Gaussian distribution are also Gaussians.

Thanks to the National Science Council of Taiwan for funding this work.

2.2. Traditional GMM-based Feature Compensation

Traditional feature compensation based on GMM trained on parallel data can be summarized as follows.

1. minimum mean squared error (MMSE)

$$\hat{\mathbf{y}}_{\text{MMSE}} = \mathbb{E}[\mathbf{y}|\mathbf{x}] = \sum_{k=1}^{K} p(k|\mathbf{x})(\mathbf{A}_k\mathbf{x} + \mathbf{b}_k), \quad (3)$$

where $\mathbb{E}$ is the expectation value operator and

$$\begin{aligned} \mathbf{A}_k &= \boldsymbol{\Sigma}_{\mathbf{yx},k}\boldsymbol{\Sigma}_{\mathbf{xx},k}^{-1}, \\ \mathbf{b}_k &= \boldsymbol{\mu}_{\mathbf{y},k} - \boldsymbol{\Sigma}_{\mathbf{yx},k}\boldsymbol{\Sigma}_{\mathbf{xx},k}^{-1}\boldsymbol{\mu}_{\mathbf{x},k}. \end{aligned} \quad (4)$$

2. stereo piecewise linear compensation for environment (SPLICE)

$$\hat{\mathbf{y}}_{\text{SPLICE}} = \sum_{k=1}^{K} p(k|\mathbf{x})(\mathbf{x} + \mathbf{r}_k), \quad (5)$$

where $\mathbf{r}_k$ is the averaged bias of mixture k over the training parallel data $\{\mathbf{x}_n, \mathbf{y}_n\}_{n=1}^{N}$ given by

$$\mathbf{r}_k = \frac{\sum_{n=1}^{N} p(k|\mathbf{x}_n)(\mathbf{x}_n - \mathbf{y}_n)}{\sum_{n=1}^{N} p(k|\mathbf{x}_n)}. \quad (6)$$

2.3. Proposed Method

In the aforementioned MMSE compensation, given in (3), the compensated feature is essentially a weighted sum of the mean vectors of the GMM for clean speech features. Each feature vector is compensated independently, so the continuity and dependence between adjacent frames are not explicitly incorporated.

In this paper, we introduce yet another Gaussian mixture model (YGMM) for feature compensation. The difference between YGMM and traditional GMM-based methods can be read from the following equation

$$\begin{aligned} \hat{\mathbf{y}}_{\text{N-MMSE}}(\mathbf{x}) &= \mathbf{x} - \boldsymbol{\beta}(\mathbf{x}) \\ &= \mathbf{x} - \sum_{k=1}^{K} p(k|\mathbf{x})\boldsymbol{\beta}_k \\ &= \mathbf{x} - \sum_{k=1}^{K} p(k|\mathbf{x})\left(\boldsymbol{\mu}_{\mathbf{x},k} - \boldsymbol{\mu}_{\mathbf{y},k}\right). \end{aligned} \quad (7)$$

The interpretation of YGMM is as follows. Instead of a weighted sum of the means for clean feature GMM, the estimation of $\hat{\mathbf{y}}$ is based on subtracting an estimated bias, denoted by $\boldsymbol{\beta}$, from the noisy speech feature $\mathbf{x}$. It is literally "noise-subtraction", and the continuity between adjacent frames is maintained in $\mathbf{x}$.

3. EVALUATION

3.1. Aurora 2.0 Database

The proposed YGMM method for feature compensation is evaluated on the Aurora 2.0 database [23]. Eight types of additive noises are artificially added to clean speech data with SNR levels ranging from 20 to -5 dB. The data may be further convolved with two types of convolution noises. There are 2 training sets called multi-train and clean-train. These sets are completely parallel and each set contains 8,440 utterances. There are 3 test sets. Test data in Set A are matched to the multi-condition train data, test data in Set B are not matched to the multi-condition train data, and test data in Set C are further mismatched due to convolution. A brief summary of database is provided in Table 1.

3.2. GMM for Classification and Compensation

The block diagram of the proposed system is illustrated in Fig. 1. The condition of a test utterance is decided by a condition classifier. Subsequently, the feature vectors are compensated accordingly if the utterance is noisy.

The basic speech feature vector consists of the 12 mel-frequency cepstral coefficients (MFCC) $c_1, \ldots, c_{12}$ and the log energy. For feature compensation, we concatenate each parallel (noisy, clean) feature vectors $(\mathbf{x}, \mathbf{y})$ in the training data into a 26-dimensional vector $\mathbf{z}$. There are 4 noise types and 4 noise levels in the training data, i.e.,

$$\{\text{subway, babble, car, exhibition hall}\} \times \{20, 15, 10, 5\}\,\text{dB},$$

so 16 GMMs of 256 mixtures each are trained for feature compensation. For a test utterance, the GMM used for feature compensation is determined by a condition classifier. The GMMs in the condition classifier is trained using the multi-condition training data. For classification, only the first 10 frames of an utterance is used for training, and the feature vectors are only 13-dimensional. There are 17 GMMs and each GMM contains 4 mixtures. A brief summarization is provided in Table 2.

The parameters in GMMs are initialized by a K-means clustering algorithm, which itself is initialized by randomly-chosen K data points. The GMM parameters are then re-estimated via an EM-algorithm with stopping criteria on log-likelihood difference and number of epochs.

3.3. Recognizer

A hidden Markov model-based (HMM) recognizer is used in the backend. Specifically, 16-state whole-word models are used for digits, along with a 3-state silence model, and a 1-state short-pause model. The state of the short-pause model is tied to the middle state of the silence model. The state-emitting probability density is a 3-component Gaussian mixture for a word state, and a 6-component Gaussian mixture for

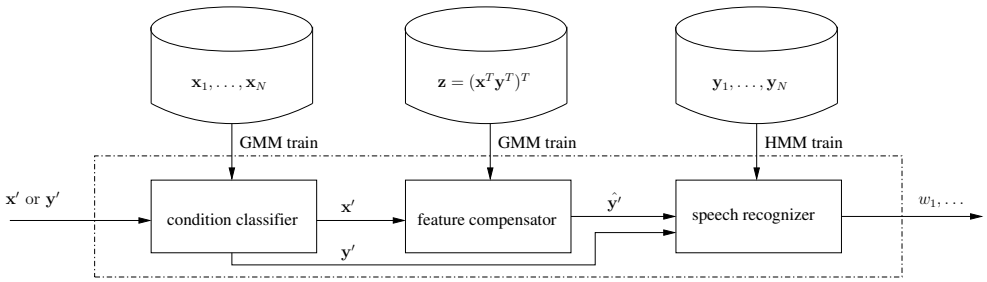

Fig. 1. Block Diagram of the Proposed System

Table 1. Aurora 2.0 Database Summary

content	strings with 1–7 English digits
background noise	subway, babble, car, exhibition hall restaurant, street, airport, train
channel noise	simulated telephone channels
noise level	20 dB, 15 dB, 10 dB, 5 dB, 0 dB, -5 dB
train set	multi-train, clean-train
test set	Set A, Set B, Set C

Table 2. Summary of Used Gaussian Mixture Models

	classification	compensation
feature dimension	13	26
mixtures per GMM	4	256
number of GMMs	17	16
noise type	subway, babble, car, exhibition hall	
noise level	20 dB, 15 dB, 10 dB, 5 dB	

a silence/short-pause state. During HMM training and recognition, the dynamic features of velocity and acceleration are also derived, resulting in a 39-dimension vector per frame. The clean-train data set, consisting of 8,440 purely clean utterances, is used to train the recognizer.

3.4. Results and Discussion

The word *accuracy* rates of the evaluation results are tabulated in Table 3. Overall, YGMM achieves a better performance than baseline and MMSE in noisy conditions (0–20 dB) and extremely noisy conditions (-5 dB) without lowering the performance in clean condition.

The relative improvements in averaged word *error* rate over 0–20 dB signal-to-noise ratio (SNR) test data can be obtained from Table 3 and they are summarized in Table 4. In this commonly-encountered context, YGMM achieves a relative improvement of 52.4% over the standard baseline (39.9% $\rightarrow$ 19.0%), and 24.9% (25.3% $\rightarrow$ 19.0%) over traditional MMSE. The results show that directly subtract estimated noise effect from the noisy feature leads to better performance.

As the training data for GMM compensation and classification are partially matched to Set A, the degree of data mismatch gets worse from Set A to Set B, and even worse for Set C. As a result, the performance of the proposed system is best in Set A (88.3% for 0–20 dB test data), and is worst in Set C (74.2%).

Table 4. Relative Improvement in Word Error Rate

	Avg	over baseline	over MMSE
baseline	39.9	=	
MMSE	25.3	36.8	=
N-MMSE	19.0	52.4	24.9

4. CONCLUSION AND FEATURE WORK

In this research, we propose a GMM-based noise-robust feature compensation method abbreviated YGMM. The basic idea of YGMM is to directly estimate noise effect on feature and subtract this effect from noisy features. On Aurora 2.0 database which is completely parallel, YGMM achieves a better performance than traditional compensation method based on MMSE.

The number of GMMs is a design parameter, as well as the number of mixtures in each GMM. With more data and more diverse test environments, these number can be increased to achieve better performance. Thus, it would be interesting to extend this work to other database, such as Aurora 3.0, to study the scalability. Indeed, even cross-lingual schemes for compensation can be reasonably attempted.

5. RELATION TO PRIOR WORK

The notation and formulation for Gaussian models follows [22]. Using GMMs to model parallel data for feature compensation have been studied in [18, 19] (MMSE) and in [20] (SPLICE). A compensation scheme taking into account the constraints between adjacent frames [21] is implemented.

Table 3. Word Accuracy Rates on Aurora 2.0 Database

Baseline														
	A					B					C			Overall
	Sub.	Bab.	Car	Exhi.	Avg	Rest.	Street	Air.	Sta.	Avg	Sub.M	Str.M	Avg	Avg
Clean	98.9	99.0	99.0	99.2	99.0	98.9	99.0	99.0	99.2	99.0	99.1	99.0	99.1	99.0
20 dB	97.1	90.2	97.4	96.4	95.3	90.0	95.7	90.6	94.7	92.8	93.5	95.1	94.3	94.1
15 dB	93.5	73.8	90.0	72.0	87.3	76.2	88.5	77.0	83.7	81.3	86.8	88.9	87.8	85.0
10 dB	78.7	49.4	67.0	75.7	67.7	54.8	67.1	53.9	60.3	59.0	73.9	74.4	74.2	65.5
5 dB	52.2	26.8	34.1	44.8	39.5	31.0	38.5	30.3	27.9	31.9	51.3	49.2	50.2	38.6
0 dB	26.0	9.3	14.5	18.1	17.0	11.0	17.8	14.4	11.6	13.7	25.4	22.9	24.2	17.1
-5 dB	11.2	1.6	9.4	9.6	7.9	3.5	10.5	8.2	8.5	7.7	11.8	11.2	11.5	8.5
Avg	69.5	49.9	60.6	65.4	61.3	52.6	61.5	53.3	55.6	55.8	66.2	66.1	66.1	60.1
GMM MMSE														
	A					B					C			Overall
	Sub.	Bab.	Car	Exhi.	Avg	Rest.	Street	Air.	Sta.	Avg	Sub.M	Str.M	Avg	Avg
Clean	98.9	99.0	99.0	99.2	99.0	98.9	99.0	98.9	99.3	99.0	99.1	98.9	99.0	99.0
20 dB	95.6	95.5	96.3	96.5	96.0	95.4	93.9	95.2	96.0	95.1	92.0	90.9	91.4	94.7
15 dB	92.3	93.8	93.9	93.4	93.3	92.6	90.0	92.2	93.5	92.1	85.4	82.7	84.0	91.0
10 dB	88.7	86.1	89.1	87.8	87.9	84.3	81.7	84.3	83.5	83.4	72.3	70.3	71.3	82.8
5 dB	80.7	67.6	75.1	79.5	75.7	63.8	59.8	64.9	66.5	63.8	46.4	50.1	48.3	65.4
0 dB	59.7	39.1	48.1	57.8	51.2	36.7	31.2	36.3	37.0	35.3	21.5	26.0	23.7	39.4
-5 dB	28.1	12.9	21.1	30.5	23.1	8.7	10.2	7.0	12.5	9.6	11.1	13.6	12.3	15.6
Avg	83.4	76.4	80.5	83.0	80.8	74.6	71.3	74.6	75.3	73.9	63.5	64.0	63.7	74.7
YGMM														
	A					B					C			Overall
	Sub.	Bab.	Car	Exhi.	Avg	Rest.	Street	Air.	Sta.	Avg	Sub.M	Str.M	Avg	Avg
Clean	98.9	99.0	99.0	99.2	99.0	98.9	99.0	99.0	99.2	99.0	99.1	99.0	99.1	99.0
20 dB	97.7	98.1	98.4	98.1	98.1	97.6	97.3	97.4	98.2	97.6	97.5	96.6	97.0	97.7
15 dB	96.2	97.2	97.6	96.4	96.8	96.2	95.3	95.4	96.6	95.9	94.7	93.2	93.9	95.9
10 dB	93.5	93.4	94.5	93.2	93.6	90.3	87.8	89.3	90.6	89.5	85.9	83.1	84.5	90.2
5 dB	87.0	78.0	83.6	85.3	83.5	72.8	68.1	72.2	74.2	71.8	63.3	63.5	63.4	74.8
0 dB	68.5	44.6	58.5	68.3	60.0	42.7	39.2	39.6	41.5	40.2	31.0	33.2	32.1	46.7
-5 dB	33.8	12.2	23.4	36.6	26.5	9.8	13.3	4.6	10.2	9.5	13.1	16.1	14.6	17.3
Avg	88.6	82.3	86.5	88.3	86.4	79.9	77.5	78.8	80.2	79.1	74.5	73.9	74.2	81.0

6. REFERENCES

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Audio, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[2] A. Berstein and I. Shallom, "An hypothesized wiener filtering approach to noisy speech recognition," *in Proceedings of the Acoustics, Speech, and Signal Processing*, vol. 2, pp. 913–916, Apr. 1991.

[3] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Audio, Speech, and Signal Processing*, vol. 29, no. 2, pp. 254–272, Apr. 1981.

[4] O. Viikki, D. Bye, and K. Laurila, "A recursive feature vector normalization approach for robust speech recognition in noise," in *proceedings of 1998 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), Seattle*, May 1998, vol. 2, pp. 733–736.

[5] A. de La Torre, A. Peinado, J. Segura, J. Perez-Cordoba, M. Benitez, and A. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 355–366, 2005.

[6] Jort F. Gemmeke, Tuomas Virtanen, and Antti Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech and Language processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[7] A. Adler, V. Emiya, M.G. Jafari, M. Elad, R. Gribonval, and M.D. Plumbley, "Audio inpainting," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 3, pp. 922–932, 2012.

[8] T. N. Sainath, B. Ramabhadran, M. Picheny, D. Nahamoo, and D. Kanevsky, "Exemplar-based sparse representation features: From TIMIT to LVCSR," *IEEE Transactions on Speech and Audio Processing*, vol. 19, no. 8, pp. 2598–2613, November 2011.

[9] D. Dimitriadis, P. Maragos, and A. Potamianos, "On the effects of filterbank design and energy computation on robust speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 6, pp. 1504–1516, 2011.

[10] L. Liu and J. He, "On the use of orthogonal GMM in speaker recognition," in *proceedings of 1999 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), Phoenix, Arizona*, 1999, vol. 2, pp. 845–848.

[11] H.R.S. Mohammadi, R. Saeidi, M.R. Rohani, and R.D. Rodman, "Combined inter-frame and intra-frame fast scoring methods for efficient implementation of GMM-based speaker verification systems," in *proceedings of 2007 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), Honolulu, Hawaii, USA*, 2007, vol. 4, pp. 309–312.

[12] Yu-Jin Kim and Jae-Ho Chung, "Signal bias removal based GMM for robust speaker recognition," in *proceedings of 2002 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), Honolulu, Hawaii, USA*, May 2002, vol. 4, p. 4163.

[13] Xiaodan Zhuang, Jing Huang, G. Potamianos, and M. Hasegawa-Johnson, "Acoustic fall detection using Gaussian mixture models and GMM supervectors," in *proceedings of 2009 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), Taipei, Taiwan*, 2009, pp. 69–72.

[14] Y. Chen, M. Chu, E. Chang, and J. Liu, "Voice conversion with smoothed GMM and MAP adaptation," in *Proceedings of INTERSPEECH. 2003*, 2003, pp. 2413–2416.

[15] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *proceedings of 1998 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), Seattle*, 1998, vol. 1, pp. 285–288.

[16] Y. Strlianou, O. Cappe, and E. Moulines, "Histogram equalization of speech representation for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 355–366, 2005.

[17] T. Toda, A.W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[18] M. Afify, X. Cui, and Y. Gao, "Stereo-based stochastic mapping for robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1325–1334, 2009.

[19] X. Cui, M. Afify, and Y. Gao, "MMSE-based stereo feature stochastic mapping for noise robust speech recognition," in *proceedings of 2008 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), Caesars Palace, Las Vegas, Nevada, USA*, 2008, pp. 4077–4080.

[20] Li Deng, A. Acero, Li Jiang, J. Droppo, and Xuedong Huang, "High-performance robust speech recognition using stereo training data," in *proceedings of 2001 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), Salt Lake City, Utah*, 2001, vol. 1, pp. 301–304.

[21] H. Zen, Y. Nankaku, and K. Tokuda, "Continuous stochastic feature mapping based on trajectory HMM s," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 2, pp. 417–430, 2011.

[22] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[23] D. Pearce and H.G. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ICSA ITRW ASR2000*, September 2000.