Revisiting Hidden Markov Models for Speech Emotion Recognition

Shuiyang Mao†, Dehua Tao, Guangyan Zhang, P. C. Ching and Tan Lee

Department of Elec. Eng., The Chinese University of Hong Kong, Hong Kong SAR, China

E-mail: †symao@ee.cuhk.edu.hk

Background & Motivation

- **Speech Emotion Recognition:**
  - Extracting the emotional state of a speaker from his or her speech;
  - In this study, we consider categorical representations (i.e., happiness, sadness, anger, etc.) for utterance-level speech emotion recognition.
- **Application:**
  - Human machine interaction (HCI);
  - Monitoring, control and psychological consultations.
- **Standard Framework:**
  - Extraction of emotion-specific features;
  - Decision making based on the extracted features.
- **Contributions:**
  - Investigate three hidden Markov model (HMM) based architectures for utterance-level speech emotion recognition;
  - Propose to improve the emotion recognition rate by incorporating various advanced techniques from the automatic speech recognition area.

The HMM based Architectures for Speech Emotion Recognition

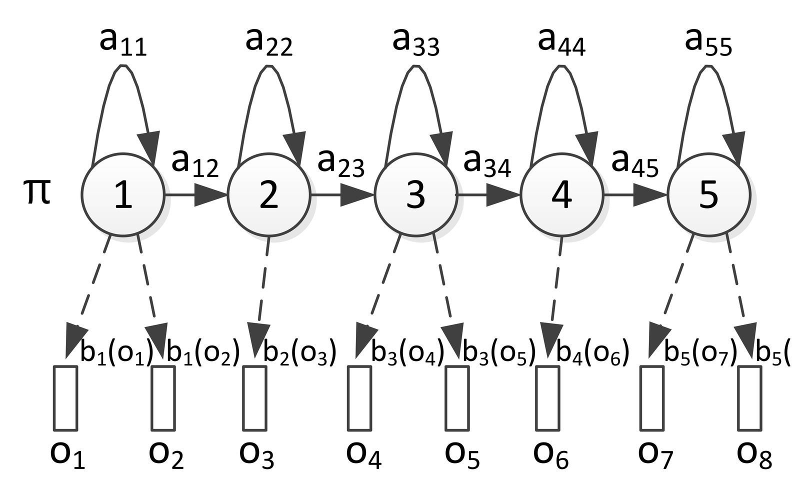

Figure 1: The Hidden Markov Generation Model

- **Hidden Markov Model (HMM):**
  - An HMM is a generative model in which the system being modeled is assumed to be a Markov process with hidden states (Fig. 1);
  - In this work, we develop C HMMs $\{\lambda_c, (c = 1, ..., C)\}$ for C discrete emotions, where C varies among database;
  - For an unknown input speech utterance $\mathbf{O}$, it is assigned to the emotion label

$$c^* = \underset{1\le c\le C}{\operatorname{argmax}} P(\mathbf{O}|\lambda_c) \tag{1}$$

where $P(\mathbf{O}|\lambda_c)$ is calculated using the Viterbi algorithm.

- **GMM-HMM Based Speech Emotion Recognition:**
  - In GMM-HMM, the observation function for the HMM state s_i is defined as a weighted sum of M_i multivariate Gaussian functions:

$$b_i(\mathbf{o}_t) = P(\mathbf{o}_t|q_t = s_i) = \sum_{l=1}^{M_i} \omega_{il}\mathcal{N}(\mathbf{o}_t|\boldsymbol{\mu}_{il}, \boldsymbol{\Sigma}_{il}) \tag{2}$$

where $\mathcal{N}(\mathbf{o}_t|\boldsymbol{\mu}_{il}, \boldsymbol{\Sigma}_{il})$ is a Gaussian component with mean vector $\boldsymbol{\mu}_{il}$ and covariance matrix $\boldsymbol{\Sigma}_{il}$. For a feature vector $\mathbf{o}_t$ of dimension n:

$$\mathcal{N}(\mathbf{o}_t|\boldsymbol{\mu}_{il}, \boldsymbol{\Sigma}_{il}) = \frac{exp\{-\frac{1}{2}(\mathbf{o}_t - \boldsymbol{\mu}_{il})^T\boldsymbol{\Sigma}_{il}^{-1}(\mathbf{o}_t - \boldsymbol{\mu}_{il})\}}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}_{il}|}} \tag{3}$$

ω_{il} denotes the mixture weight of Gaussian component l of state s_i, and the weights are subject to $\sum_{l=1}^{M_i}\omega_{il} = 1$.

- **SGMM-HMM Based Speech Emotion Recognition:**
  - **Drawbacks of GMM-HMM:** Involves training a completely separate GMM in each HMM state, which might suffer from over-fitting;
  - In SGMM-HMM, the covariance matrix for each GMM component is shared between states, whereas the mean and mixture weights are allowed to vary in a subspace of the full parameter space, thus providing a more compact model representation;
  - The observation function for a SGMM-HMM at some state s_i has the following form:

$$b_i(\mathbf{o}_t) = P(\mathbf{o}_t|q_t = s_i) = \sum_{l=1}^{M} \omega_{il}\mathcal{N}(\mathbf{o}_t|\boldsymbol{\mu}_{il}, \boldsymbol{\Sigma}_l) \tag{4}$$

where $\boldsymbol{\mu}_{il}$ is computed using linear subspace projection matrix $\mathbf{M}_l$ and projection vector $\mathbf{v}_i$ for the state s_i:

$$\boldsymbol{\mu}_{il} = \mathbf{m}_l + \mathbf{M}_l\mathbf{v}_i \tag{5}$$

and the mixture weight ω_{il} is computed from linear subspace projection vector $\mathbf{w}_l$ and the same sate-dependent projection vector $\mathbf{v}_i$:

$$\omega_{il} = \frac{exp\{\mathbf{w}_l^T\mathbf{v}_i\}}{\sum_{j=1}^{M} exp\{\mathbf{w}_j^T\mathbf{v}_i\}} \tag{6}$$

- **DNN-HMM Based Speech Emotion Recognition:**
  - **Drawbacks of GMM-HMM and SGMM-HMM:** Statistically inefficient to model non-linear data in the feature space;
  - In DNN-HMM, the GMMs (or SGMMs) are replaced with DNN to estimate the observation probabilities of input acoustic features at each HMM state;
  - All of the training utterances, combined with their labeled state sequence which are generated from GMM-HMM or SGMM-HMM alignment, are fed as inputs to train the DNN;
  - The outputs of the DNN are the posterior probabilities of the $C \times Q$ output units, with C and Q denoting the emotion class number and HMM state number, respectively;
  - According to the Bayesian theorem, the observation probability $p(\mathbf{o}_t|q_t)$ is calculated as follows:

$$p(\mathbf{o}_t|q_t) = \frac{p(q_t|\mathbf{o}_t)p(\mathbf{o}_t)}{p(q_t)} \tag{7}$$

where $p(q_t)$ is estimated from an initial state-level alignment of the training set; and $p(\mathbf{o}_t)$ is independent of the state sequence, and thus can be ignored.

Speech Corpora

- Three corpora of acted emotions are used to evaluate the validity and universality of our approach: a Chinese emotional corpus (CASIA), a German emotional corpus (Emo-DB), and an English emotional database (IEMOCAP), which are summarized in Fig. 2.

Corpora	Language	#Utterance	#Subjects	#Emotion
CASIA	Chinese	7,200	4 (2 female)	6
Emo-DB	German	420	10 (5 female)	5
IEMOCAP	English	5,347	10 (5 female)	4

Figure 2: Overview of the selected emotion corpora. (#Utterance: number of utterances used, #Subjects: number of subjects, and #Emotion: number of emotions involved.)

	Speaker-dependent						Speaker-independent					
	CASIA		Emo-DB		IEMOCAP		CASIA		Emo-DB		IEMOCAP	
	UA [%]	WA [%]	UA [%]	WA [%]	UA [%]	WA [%]	UA [%]	WA [%]	UA [%]	WA [%]	UA [%]	WA [%]
(1) GMM-HMM	76.60	76.60	77.45	82.14	61.59	59.59	44.31	44.31	85.02	86.43	57.65	53.00
(2) GMM-HMM(ST)	79.93	79.93	81.15	83.33	63.51	61.93	46.33	46.33	86.15	87.38	59.54	53.80
(3) GMM-HMM(ST+SAT)	83.26	83.26	83.95	85.71	64.33	63.33	50.44	50.44	85.50	87.38	60.25	55.00
(4) SGMM-HMM	86.88	86.88	**88.25**	**90.48**	66.63	64.83	**53.81**	**53.81**	**86.23**	**87.62**	61.77	56.40
(5) SGMM-HMM(MMI)	87.50	87.50	–	–	**66.94**	**65.86**	52.69	52.69	–	–	**62.23**	57.20
(6) DNN-HMM(GMM-Ali.)	90.74	90.74	64.38	69.56	65.20	64.66	38.35	38.35	64.69	65.28	57.12	60.13
(7) DNN-HMM(SGMM-Ali.)	**91.32**	**91.32**	64.60	71.43	65.12	64.17	39.40	39.40	64.71	67.38	58.02	**62.28**

Figure 3: Comparison of unweighted accuracy and weighted accuracy on different HMM based architectures on CASIA corpus, Emo-DB corpus and IEMOCAP database, respectively. (ST: HMM state tying, SAT: speaker adaptive training, MMI: sequential discriminative training with maximum mutual information criterion, GMM (SGMM) -Ali.: alignment generated from monophone GMM-HMM (SGMM-HMM))

Experimental Settings

- **Acoustic Features:**
  - 15-dimensional MFCCs with the first- and second-order derivatives + pitch + voicing probability.
- **DNN Architecture:**
  - One input layer, three hidden layers with 256 neurons per layer, followed by one softmax loss layer;
  - A hyperbolic tangent non-linearity is applied between two consecutive hidden layers.
- **DNN Training:**
  - Frame classification training is based on mini-batch Stochastic Gradient Descent, optimizing frame cross-entropy;
  - The initial learning rate of 0.015 is gradually decreased to 0.002 after 20 epochs.
- **Both speaker-dependent (SD) and speaker-independent (SI) scenarios are considered:**
  - SD: Randomly select 80% as the training set, 10% as the validation set and the rest 10% as the test set;
  - SI: K-folds leave-one-speaker-out cross-validation, where K denotes the number of speakers in each database.

Results & Analysis

- Fig. 3 shows the performance comparison between different HMM based systems on three corpora;
- Comparison of recognition accuracy on CASIA. (Spk-Dep.: speaker-dependent, and Spk-Indep.: speaker-independent.)

Methods for comparison	Spk-Dep. [%]	Spk-Indep. [%]
Sun et al. [1] (2015)	85.08	43.50
Wen et al. [2] (2017)	–	48.50
Liu et al. [3] (2018)	90.28	38.55
Our method		
GMM-HMM(ST+SAT)	83.26	50.44
SGMM-HMM	86.88	**53.81**
DNN-HMM(SGMM-Ali.)	**91.32**	39.40

- Comparison of weighted accuracy on Emo-DB for speaker-independent task. (#Emotion: number of emotions used in each experiment.)

Methods for comparison	#Emotion	W. Accuracy [%]
Li et al. [4] (2016)	4	86.38
Semwal et al. [5] (2017)	6	80.00
Wen et al. [2] (2017)	7	82.32
Our method		
GMM-HMM(ST+SAT)	5	87.38
SGMM-HMM	5	**87.62**
DNN-HMM(SGMM-Ali.)	5	67.38

- Comparison of unweighted accuracy and weighted accuracy on IEMOCAP for speaker-independent task.

Methods for Comparison	U. Accuracy [%]	W. Accuracy [%]
Huang et al. [6] (2016)	49.96	59.33
Ma et al. [7] (2017)	62.54	57.85
Mirsamadi et al. [8] (2017)	58.80	**63.50**
Luo et al. [9] (2018)	**63.98**	60.35
Our Method		
GMM-HMM(ST+SAT)	60.25	55.00
SGMM-HMM(MMI)	62.23	57.20
DNN-HMM(SGMM-Ali.)	58.02	62.28

References

[1] Y. Sun, G. Wen, and J. Wang, "Weighted spectral features based on local hu moments for speech emotion recognition," *Biomed. Signal Process. Control*, vol. 18, pp. 80–90, 2015.

[2] G. Wen, H. Li, J. Huang, D. Li, and E. Xun, "Random deep belief networks for recognizing emotions from speech signals," *Comput. Intell. Neurosci.*, vol. 2017, no. 2, pp. 1–9, 2017.

[3] Z. Liu, Q. Xie, M. Wu, W. Cao, Y. Mei, and J. Mao, "Speech emotion recognition based on an improved brain emotion learning model," *Neurocomputing*, vol. 309, pp. 145–156, 2018.

[4] X. Li and M. Akagi, "Multilingual speech emotion recognition system based on a three-layer model." in *Proc. INTERSPEECH*, 2016, pp. 3608–3612.

[5] N. Semwal, A. Kumar, and S. Narayanan, "Automatic speech emotion detection system using multi-domain acoustic feature selection and classification models," in *Proc. ISBA*, 2017, pp. 1–6.

[6] C. Huang and S. S. Narayanan, "Attention assisted discovery of sub-utterance structure in speech emotion recognition." in *Proc. INTERSPEECH*, 2016, pp. 1387–1391.

[7] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng, and L. Cai, "Speech emotion recognition with emotion-pair based framework considering emotion distribution information in dimensional emotion space," in *Proc. INTERSPEECH*, 2017, pp. 1238–1242.

[8] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. ICASSP*, 2017, pp. 2227–2231.

[9] D. Luo, Y. Zou, and D. Huang, "Investigation on joint representation learning for robust feature extraction in speech emotion recognition," in *Proc. INTERSPEECH*, 2018, pp. 152–156.