# REVISITING HIDDEN MARKOV MODELS FOR SPEECH EMOTION RECOGNITION

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

Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China

**ABSTRACT**

Hidden Markov models (HMMs) have a long tradition in automatic speech recognition (ASR) due to their capability of capturing temporal dynamic characteristics of speech. For emotion recognition from speech, three HMM based architectures are investigated and compared throughout the current paper, namely, the Gaussian mixture model based HMMs (GMM-HMMs), the subspace based Gaussian mixture model based HMMs (SGMM-HMMs) and the hybrid deep neural network HMMs (DNN-HMMs). Extensive emotion recognition experiments are carried out on these three architectures on the CASIA corpus, the Emo-DB corpus and the IEMOCAP database, respectively, and results are compared with those of state-of-the-art approaches. These HMM based architectures prove capable of constituting an effective model for speech emotion recogntion. Also, the modeling accuracy is further enhanced by incorporating various advanced techniques from the ASR area. In particular, among all of the architectures, the SGMM-HMMs achieve the best performance in most of the experiments.

***Index Terms***— Speech emotion recognition, hidden Markov models, subspace based GMM, hybrid DNN-HMM

## 1. INTRODUCTION

Speech emotion recognition aims at detecting the underlying emotional state of the speaker from his or her speech. Especially in the field of human-machine interaction (HCI), growing interest can be observed in recent years. In addition, the detection of lies, monitoring of call centres and psychological consultations are often claimed as promising application scenarios for speech emotion recognition.

In emotion classification of speech signals, the popular features employed are statistics of fundamental frequency (pitch), spectral shape and energy contour. These statistical measures are generally estimated over the whole utterance, and thus termed global features. Several studies find a high correlation between some statistics of speech and the emotional state of the speaker [1–3]. However, using global features also presents several drawbacks: first, as is well known, emotional information conveyed by speech is inherently sequential, while taking global statistics ignores such temporal behaviour; second, the performance of systems employing these global features generally degrades substantially when more than two categories of emotion are to be classified [4]; and third, the recognition process can only be performed once the whole utterance has been pronounced, which limits the capability of building real time recognisers.

A different approach to global statistics is to use frame based raw features such as Mel Frequency Cepstrum Coefficients (MFCCs), energy or pitch. For the purpose of explicitly performing temporal modeling based on these raw features in order to better exploit the dynamic information of emotional speech, dynamic models, such as hidden Markov models (HMMs) [5–8], are frequently considered. HMMs have formed the core of statistical ASR for over three decades. The underlying idea is that speech signals are not stationary and can be modeled as a concatenation of HMM states, with each modeling different sounds or sound combinations and having their own statistical properties. In addition to HMMs, recurrent neural networks (RNNs), e.g., with long short-memory (LSTM) [9], are also effective at dynamic modeling, and a growing trend exists to apply LSTM based architectures for speech emotion recognition [10–13]. The main advantage of HMMs over these emerging LSTMs for speech emotion recognition is that HMM based architectures have long been studied in the ASR area, comprising available and well established procedures for optimizing the recognition framework, e.g., Viterbi decoding, sequential discriminative training, speaker adaptive training, etc.

In this work, we extend beyond feature selection and explore three HMM based architectures for text-independent speech emotion recognition. The first one is the classic Gaussian mixture model based HMMs (GMM-HMMs), in which GMMs serve as state observing functions. For the last few decades, GMMs have been the most widely-utilized density functions for likelihood computation in ASR. However, conventional GMMs in HMM possess several major shortcomings. For instance, they generally involve training a completely separate GMM in each HMM state, which may suffer from an over-fitting problem in applications, such as speech emotion recognition, where there is usually a small amount of (in-domain) training data. To deal with this over-fitting issue, we investigate the subspace based Gaussian mixture model based HMMs (SGMM-HMMs), in which the HMM states share a common structure, whereas the mean and mixture weights are allowed to vary in a subspace of the full parameter space, controlled by the projection vectors of low dimension. This leads to a significant decrease in the total number of parameters, which we conjecture might offer an advantage in emotion recognition tasks. To the best of our knowledge, this is the first attempt to investigate SGMM-HMM for speech emotion recognition.

Another limitation of GMMs (or SGMMs) in HMM is that they are statistically inefficient for modeling data that lie on or near a nonlinear manifold in the data space [14]. Therefore, we replace the GMMs (or SGMMs) with a deep neural network (DNN) to estimate the (scaled) observation likelihood, which forms the hybrid deep neural network HMM (DNN-HMM) structure. The DNN is capable of capturing the underlying nonlinear relationship among data, and can be viewed as a powerful discriminative feature extractor mining high-level representations optimized to predict the emotion class [15].

The remainder of this paper is organized as follows. In Section 2, we briefly describe three HMM based architectures, and present basic concepts of how they are applied to speech emotion recognition tasks. In Section 3, we briefly describe three popular emotional corpora on which extensive experiments are conducted. In Section 4 experimental results using these three emotional corpora are discussed. Finally, Section 5 draws conclusions and outlines directions for future work.

## 2. HMM BASED ARCHITECTURE FOR SPEECH EMOTION RECOGNITION

Hidden Markov models (HMMs) have been successfully applied as a core for statistical acoustic models in many systems. An HMM is a generative model in which the system being modeled is assumed to be a Markov process with hidden states (such as emotion-dependent states, hidden in the speech signal). A typical HMM, represented as $\lambda = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$, consists of the following elements:

1. State sequence $S = \{s_1, s_2, ..., s_Q\}$, where $Q$ is denoted as the number of states in each HMM model, usually from 3 to 5, and $q_t \in \{s_1, s_2, ..., s_Q\}$ is the HMM state at time $t$.
2. Transition probability matrix $\mathbf{A} = \{a_{ij}\}$, with $a_{ij} = P(q_t = s_j | q_{t-1} = s_i)$, $1 \leq i, j \leq Q$.
3. Observation functions $\mathbf{B} = \{b_i(\mathbf{o_t})\}$, where $b_i(\mathbf{o_t})$ represents the probability of observing $\boldsymbol{o}_t$ at state $s_i$, $1 \leq i \leq Q$.
4. Initial state distribution $\boldsymbol{\pi} = \{\pi_i\}$, where $\pi_i = P(q_1 = s_i)$, $1 \leq i \leq Q$.

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

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

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

### 2.1. 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 $n$:

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

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

### 2.2. SGMM-HMM Based Speech Emotion Recognition

The GMM-HMM based framework generally involves training a completely separate GMM in each HMM state, which might suffer from over-fitting, especially for tasks such as speech emoition recognition, in which the amount of (in-domain) data available to train the model is often limited. To mitigate this problem, we introduce the subspace based GMM-HMM (SGMM-HMM), which was originally motivated by subspace-based speaker adaptation and speaker verification approaches [16]. In SGMM-HMM the HMM states share a common structure, whereas the mean and mixture weights are allowed to vary in a subspace of the full parameter space, controlled by the projection vectors of low dimension, thus providing a more compact model representation.

SGMM uses the state-independent covariance matrix of the universal background model (UBM), and computes state-dependent means by linearly projecting the means of UBM. Let us define UBM as a GMM with means and covariances $\{\mathbf{m}_l, \boldsymbol{\Sigma}_l\}$, with $l$ denoting the GMM component number. Compared to the GMM density computation (Equation 2), the observation function for a SGMM-HMM at some state $s_i$ has the following form:

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

It is worth noting that Equation 4 differs from Equation 2 in that the covariance matrix for each GMM component is shared between states in SGMM-HMM, whereas it is state-dependent in GMM-HMM.

The SGMM mean vector $\boldsymbol{\mu}_{il}$ in Equation 4 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 weights are computed from linear subspace projection vector $\mathbf{w}_l$ and the same sate-dependent projection vector $\mathbf{v}_i$:

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

### 2.3. DNN-HMM Based Speech Emotion Recognition

The GMM-HMMs or SGMM-HMMs are statistically inefficient to model non-linear data in the feature space. We overcome this restriction by using deep neural networks, which can model complex and non-linear relationships between low-level acoustic features. Based on this, we have developed a hybrid DNN-HMM based architecture for speech emotion recognition, in which the GMMs (or SGMMs) are replaced with DNN to estimate the observation probabilities of input acoustic features at each HMM state. If DNNs can be trained to better predict emotional HMM states, then the DNN-HMM based model can achieve better recognition accuracy than GMM-HMMs or SGMM-HMMs.

DNN-HMMs can be trained using the embedded Viterbi algorithm. Specifically, in our implementation, we firstly train a Left-Right GMM-HMM model with five states for each emotion class using the training utterance. Then, for each utterance in the training set, the Viterbi algorithm is performed to obtain an optimal state sequence, and each state is assigned a label according to a state-label mapping table. All of the training utterances, combined with their labeled state sequence, are then fed as inputs to train a DNN using the mini-batch based gradient descent method. 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. Then 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.

## 3. SPEECH CORPORA

We use three corpora of acted emotions 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 Table 1.

More specifically, the CASIA corpus contains 9,600 utterances that are simulated by four subjects (two males and two females) in six different emotional states, namely, anger, fear, happiness, neutral, sadness and surprise. In our experiments, we only use 7,200 utterances that correspond to 300 linguistically neutral sentences with the same statements. All of the categories of emotions are selected.

The Emo-DB corpus was produced by 10 professional actors (five males and five females) in German to simulate seven different emotions. The number of spoken utterances for these seven emotions in the Berlin Emo-DB is not equally distributed: 127 anger, 81 boredom, 46 disgust, 69 fear, 71 joy, 79 neutral, and 62 sadness. In our experiments, five categories of emotions are selected, i.e., anger, boredom, joy, sadness and neutral. Thus, a total number of 420 utterances are used in this study.

The IEMOCAP database was collected using motion capture and audio/video recording over five dyadic sessions with 10 subjects. At least three evaluators annotated each utterance in the database with the categorical emotion labels chosen from the set: anger, disgust, excitement, fear, frustration, happiness, neutral, sadness, surprise and other. We consider only the utterances with majority agreement (i.e., at least two out of three evaluators assigned the same emotion label) over the emotion classes of: anger, happiness, neutral and sadness.

**Table 1**. Overview of the selected emotion corpora. (#Utterance: number of utterances used, #Subjects: number of subjects, and #Emotion: number of emotions used.)

| | 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 |

## 4. EXPERIMENTS AND RESULTS

### 4.1. Experimental Settings

In our experiments, each input signal is converted into frames by using a 25 ms window sliding at 10 ms each time. 15 MFCCs (with C0 replaced with energy) are then extracted from each frame, with a Hamming window to minimize frequency leakage. Cepstral mean variance normalization (CMVN) is performed at the utterance level to mitigate recording variations. We have also taken the first- and second-order derivatives of the normalized MFCCs. In addition to MFCCs, which can be considered as strong correlates of the vocal tract, pitch and voicing probability are also extracted to reflect vocal cord variations, providing complementary information for recognizing emotions. Therefore, the dimension of features for each frame is 47 (i.e., $15 \times 3 + 1 + 1$). Compared to many of the state-of-the-art approaches involving large feature set construction, our feature set is simple and straightforward.

For the DNN-HMM hybrid, the DNN architecture consists of one input layer, three hidden layers, followed by one softmax loss layer. Network configuration (ordered from input to output) is set to $\{47, 256, 256, 256, N\}$, where '47' and '$N$' correspond to the dimension of the input features and the total number of state labels, respectively. A hyperbolic tangent non-linearity is applied between two consecutive hidden layers. Frame classification training is based on mini-batch Stochastic Gradient Descent (SGD), optimizing frame cross-entropy. The initial learning rate of 0.015 is gradually decreased to 0.002 after 20 epochs. This DNN configuration was found to be optimal after experimenting with different sized configurations.

Our experiments are implemented on Kaldi [17], which is an open source ASR toolkit. We find that it is relatively easy to develop these HMM based architectures for speech emotion recognition by adapting the existing Kaldi recipes, based on which we can make use of several advanced technologies from the ASR area, e.g., HMM state tying, speaker adaptive training (SAT) and sequential discriminative training (SDT).

All of the three corpora do not split training data and testing data in advance, so that two experimental strategies are used. They are speaker-dependent (SD) and speaker-independent (SI). For the speaker-dependent strategy, in each database, we randomly select $80\%$ of the speech sentences as the training set, $10\%$ as the developing set to identify the optimal parameters and the rest $10\%$ as the testing set. In the speaker-independent strategy, the $K$-folds leave-one-speaker-out cross-validation method is carried out, where $K$ denotes the number of speakers in each database. For each fold, the utterances from one speaker are used as the testing set, and the utterances from the other speakers are used as the training set.

### 4.2. Results and Analysis

We first developed three GMM-HMM systems using: (1) monophone training; (2) monophone training with state tying based on the data-driven decision tree; and (3) System (2) with speaker adaptive training applied using feature space maximum likelihood linear regression (fMLLR) transformation. In our study, speaker adaptive training is applied per speaker to adapt the emotion variation of different speakers. Based on these monophone GMM-HMMs, two SGMM-HMM systems were then developed: (4) SGMM-HMM based on System (3); and (5) SGMM-HMM system with sequential discriminative training (SDT) applied using maximum mutual information (MMI) criterion. We also built two systems using the DNN-HMM hybrid architecture: (6) DNN-HMM using an alignment generated from the tied-state monophone GMM-HMM; and (7) DNN-HMM using an alignment generated from SGMM-HMM. Extensive recognition experiments were conducted on the above (1)-(7) systems, as shown in Table 1. Both weighted accuracy (WA) and unweighted accuracy (UA) are used for perfomance evaluation. Weighted accuracy is the total number of correctly classified testing samples of all classes averaged by the total number of testing samples, and unweighted accuracy is the sum of all class accuracies divided by the number of classes, without considering the number of instances per class, which better reflects overall accuracy in the presence of an imbalanced class.

Table 2 summarizes the performance comparison between different HMM-based systems on three corpora. As can been seen: (1) SGMM-HMM based systems achieved the best results in most of the experiments. This is because SGMM can provide more compact representation than the other two observation functions and thus mitigate the over-fitting problem caused by the limited amount of emotional data. It is worth noting that for the smallest Emo-DB database, SGMM-HMM obtained the best results in all experiments, which further demonstrates the effectiveness of SGMM-HMM when only limited training data are available. (2) For some of the experiments that involve the CASIA corpus and the IEMOCAP database, hybrid DNN-HMM achieved the highest recognition rates, especially for the speaker-dependent task on the CASIA corpus, where DNN-HMM significantly outperforms the other two HMM architectures. This is mainly due to the discriminative power of the deeply learned features introduced by DNN. (3) However, for the speaker-independent task on the same CASIA corpus, the performance of using DNN-HMM degraded by a large margin. This might be attributable to the fact that, in this experiment, recognition rates from

**Table 2**. Comparison of UAs and WAs 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-Ali.: alignment generated from monophone GMM-HMM, and SGMM-Ali.: alignment generated from SGMM-HMM.)

| | 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** |

both GMM-HMM and SGMM-HMM systems were very low, and the alignments generated from both systems were quite poor. Using such un-reliable state-level alignments as training labels for DNN can impart great harm to the DNN-HMM system. (4) On the other hand, using a better alignment to generate training labels for the DNN can improve accuracy. This is confirmed by the observation that DNN-HMM using alignments generated from SGMM-HMM generally achieved better results than those of DNN-HMM using alignments generated from GMM-HMM. (5) Using advanced technologies from ASR, i.e., HMM state tying, speaker adaptive training, and sequential discriminative training (MMI), greatly boosted performance, which constitutes an enormous advantage of HMM based architectures over other methods.

**Table 3**. Comparison of recognition accuracy on CASIA. (SD: speaker-dependent, and SI: speaker-independent.)

| **Methods for comparison** | **SD** [%] | **SI** [%] |
|---|---|---|
| Sun et al. [18] (2015) | 85.08 | 43.50 |
| Wen et al. [19] (2017) | – | 48.50 |
| Liu et al. [20] (2018) | 89.60 | – |
| Liu et al. [21] (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 |

**Table 4**. Comparison of WAs on Emo-DB for SI task. (#Emo.: number of emotions used in each experiment.)

| **Methods for comparison** | **#Emo.** | **WA** [%] |
|---|---|---|
| Li et al. [22] (2016) | 4 | 86.38 |
| Zhu et al. [23] (2017) | 5 | 74.12 |
| Semwal et al. [24] (2017) | 6 | 80.00 |
| Wen et al. [19] (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 |

Tables 3-5 compare our result with prior work on three corpora, respectively. For the CASIA corpus, our DNN-HMM hybrid architecture achieved the highest recognition accuracy of 91.32% on the SD task, while SGMM-HMM surpassed other methods on the SI task. It is worth noting that the gap of recognition rates between SD and SI tasks on CASIA is much larger than that of the other two corpora. This is mainly because there is only four speakers in CASIA (see Table 1). Hence for each fold, we only have training utterances from three speakers, which limits the generalization capability to utterances from an unseen speaker. Table 4 compares weighted accuracy (WA) on the Emo-DB corpus for the SI task. It is worth noting that the systems differ in number and type of emotions. Nevertheless, it provides a basic comparison of the different approaches. For the IEMOCAP corpus, which might be the most challenging dataset, our HMM based architecture also achieved a comparable result with that of state-of-the-art approaches for SI task, as shown in Table 5. As mentioned previously, our study only used a simple feature set consisting of MFCCs, pitch and voicing probability, with a dimension of less than 50. On the other hand, other state-of-the-art approaches generally used a much more complex feature set, i.e., [26] used the INTERSPEECH 2009 Emotion Challenge feature set with a dimension of 384 for recognizing emotions on the IEMOCAP corpus. This further demonstrates the effectiveness of the HMM based architectures used in our study.

**Table 5**. Comparison of UAs and WAs on IEMOCAP for SI task.

| **Methods for Comparison** | **UA** [%] | **WA** [%] |
|---|---|---|
| Han et al. [25] (2014) | 48.20 | 54.30 |
| Huang et al. [11] (2016) | 49.96 | 59.33 |
| Ma et al. [26] (2017) | 62.54 | 57.85 |
| Mirsamadi et al. [12] (2017) | 58.80 | **63.50** |
| Luo et al. [27] (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 |

## 5. CONCLUSIONS

We believe that this contribution presents important results concerning speech emotion recognition with hidden Markov model based architectures, namely, GMM-HMMs, SGMM-HMMs, and DNN-HMMs. Extensive experiments were carried out on these HMMs on the CASIA corpus, the Emo-DB corpus and the IEMOCAP database, respectively, and showed comparable results with those of state-of-the-art approaches. Thus, HMM is proven not to be outdated, but instead a considerably effective method for automatic recognition of human emotions in speech. Since we only used a simple feature set in this study, in our future work we aim to perform more fundamental research on acoustic features, which should be more robust to environment and speaker variations. More advanced DNN architectures will also be investigated in the near future.

## 6. REFERENCES


[1] B. Heuft, T. Portele, and M. Rauth, "Emotions in time domain synthesis," in *Proc. ICSLP, 1996*, pp. 1974–1977.

[2] N. Amir and S. Ron, "Towards an automatic classification of emotions in speech," in *Proc. ICSLP, 1998*, pp. 699–702.

[3] A. Iida, N. Campbell, S. Iga, F. Higuchi, and M. Yasumura, "Acoustic nature and perceptual testing of corpora of emotional speech," in *Proc. ICSLP, 1998*, pp. 1559–1562.

[4] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003.

[5] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[6] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in *Proc. ICASSP*, 2003, vol. 2, pp. II–1.

[7] B. Vlasenko, D. Prylipko, R. Böck, and A. Wendemuth, "Modeling phonetic pattern variability in favor of the creation of robust emotion classifiers for real-life applications," *Computer Speech & Language*, vol. 28, pp. 483–500, 2014.

[8] B. Vlasenko, "Emotion recognition within spoken dialog systems," *PhD thesis. University of Magdeburg*, 2011.

[9] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in NIPS*, 2014, pp. 3104–3112.

[10] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Proc. INTERSPEECH*, 2015.

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

[12] 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.

[13] W. Han, H. Ruan, X. Chen, Z. Wang, H. Li, and B. Schuller, "Towards temporal modelling of categorical speech emotion recognition," in *Proc. INTERSPEECH*, 2018, pp. 932–936.

[14] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[15] L. Li, Y. Zhao, D. Jiang, Y. Zhang, F. Wang, I. Gonzalez, E. Valentin, and H. Sahli, "Hybrid deep neural network–hidden Markov model (dnn-hmm) based speech emotion recognition," in *Proc. ACII*, 2013, pp. 312–317.

[16] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiát, A. Rastrow, et al., "Subspace Gaussian mixture models for speech recognition," in *Proc. ICASSP*, 2010, pp. 4330–4333.

[17] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., "The Kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.

[18] 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.

[19] 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.

[20] Z. Liu, M. Wu, W. Cao, J. Mao, J. Xu, and G. Tan, "Speech emotion recognition based on feature selection and extreme learning machine decision tree," *Neurocomputing*, vol. 273, pp. 271–280, 2018.

[21] 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.

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

[23] B. Zhu, W. Zhou, Y. Wang, H. Wang, and J. Cai, "End-to-end speech emotion recognition based on neural network," in *Proc. ICCT*, 2017, pp. 1634–1638.

[24] 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.

[25] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. INTERSPEECH*, 2014.

[26] 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.

[27] 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.