Speech Emotion Recognition Based on Acoustic Segment Model

Siyuan Zheng[1], *Jun Du*[*1] , *Hengshun Zhou*[1], *Xue Bai*[1], *Chin-Hui Lee*[2], *Shipeng Li*[3]

[1]National Engineering Laboratory for Speech and Language Information Processing University of Science and Technology of China
[2]Georgia Institute of Technology
[3]Shenzhen Institute of Artificial Intelligence and Robotics for Society

zsy19@mail.ustc.edu.cn, ✉jundu@ustc.edu.cn, zhhs@mail.ustc.edu.cn, byxue@mail.ustc.edu.cn, chl@ece.gathech.edu, lishipeng@cuhk.edu.cn

2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP) | 978-1-7281-6994-1/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISCSLP49672.2021.9362119

Abstract

Accurate detection of emotion from speech is a challenging task due to the variability in speech and emotion. In this paper, we propose a speech emotion recognition (SER) method based on acoustic segment model (ASM) to deal with this issue. Specifically, speech with different emotions is segmented more finely by ASM. Each of these acoustic segments is modeled by Hidden Markov Models (HMMs) and decoded into a series of ASM sequences in an unsupervised way. Then feature vectors are obtained from these sequences above by latent semantic analysis (LSA). Finally, these feature vectors are fed to a classifier. Validated on the IEMOCAP corpus, results demonstrate the proposed method outperforms the state-of-the-art methods with a weighted accuracy of 73.9% and an unweighted accuracy of 70.8% respectively.

Index Terms: speech emotion recognition, acoustic segment model, latent semantic analysis

1. Introduction

Speech emotion recognition(SER) is the task to identify which emotion is contained in human speech. It plays an important role in speech-based human-computer interaction[1] and many applications are derived by this technology, such as quality measurement in call centers[2], intelligent service robotics and remote education.

The key of SER is to find appropriate features to represent the emotion in speech. Traditional system for SER generally includes frontend feature extraction in frame-level. Then these features are merged by statistical functions into feature extraction in utterance-level. Finally, the utterance representation is sent to the backend for classification[3, 4]. In the frontend, Gaussian Mixture Models (GMM) with Mel-Frequency Cepstral Coefficients (MFCC) models the features frame by frame. In the backend, there are many classification models that have been used on SER, and support vector machine(SVM) is the most popular classifier.

Recently deep learning has been successfully applied to speech related tasks, such as speech recognition[5], speech enhancement[6],acoustic scene classification[7], and SER. In [8], DNN is used in the frontend to learn the acoustic features, and the extreme learning machine(ELM) is used as the backend classifier. In [9], the speech spectrogram is used as the input of the fully convolutional neural network(FCN). And the attention mechanism makes the model focus on specific time-frequency regions of input. Some latest work on SER are structural deformation based on DNN, such as modeling the SER task under the Dual-Sequence LSTM (DS-LSTM)[10], and the multi-time-scale (MTS) method [11].

Another latest method is using multimodal information in SER. Some researchers realize that audio data alone is not enough to make correct classification[10], so that multimodal information such as textual information[12, 13] and video information[14] are combined with the acoustic information to improve the accuracy of SER. However, it is possible that two utterances with the same textual content or two videos with similar countenances can contain entirely different meanings with different emotions. Therefore, using other multimodal information too liberally may make wrong classification on this task.

As a reason, we focus on the unimodal SER task. The acoustic information in the speech has more potential can be used for the task. Our paper proposes the acoustic segment model(ASM) framework to mining the acoustic information for the SER task. ASM is first proposed in automatic speech recognition to represent basic acoustic units and lexicons[15]. Inspired by the work, ASM has also been used in spoken language recognition[16], music retrieval[17] and acoustic scene classification[18]. As the language consists of different phonemes and grammars and the acoustic scene consists of acoustic events, the speech containing different emotions consists of fundamental units and these units are interrelated. Therefore, we can generate a sequence of acoustic units from the speech to divide different emotional dialogues.

There are many different segmentation methods to get the acoustic units on the utterances, such as even segmentation[19], maximum likelihood segmentation[20], finding the spectral discontinuities[21] and using watershed transform over the blurred self similaritydotplot[22]. In this paper, inspired by previous work in acoustic scene classification[19], each emotion is modeled by GMM-HMMs[23]. Then the speech is divided into a variable-length segment which is defined by the number of hidden states. The hidden state in each topology is corresponding to a GMM. Then the adjacent similar frames are merged to the same GMM. The sequence of GMMs is corresponding to initial acoustic model of ASM units in the speech. These ASM units are used for the audios as the initial label sequences. The ASM units are generated by the GMMs of different audios without any prior knowledge so that this approach is unsupervised learning. Then each ASM sequence is modeled by a GMM-HMM and decoded iteratively into a new sequence of ASM units. In order to extract feature vectors from these sequences, the acoustic units are regarded as terms in text documents. Then we use latent semantic analysis(LSA)[24] to generate the term-document matrix. Each column represents a feature vector of a recording. Finally, the vectors of training dataset are fed to the backend classifier such as DNN.

The remainder of the paper is organized as follows. In Section 2, the proposed architecture of our method is introduced. In

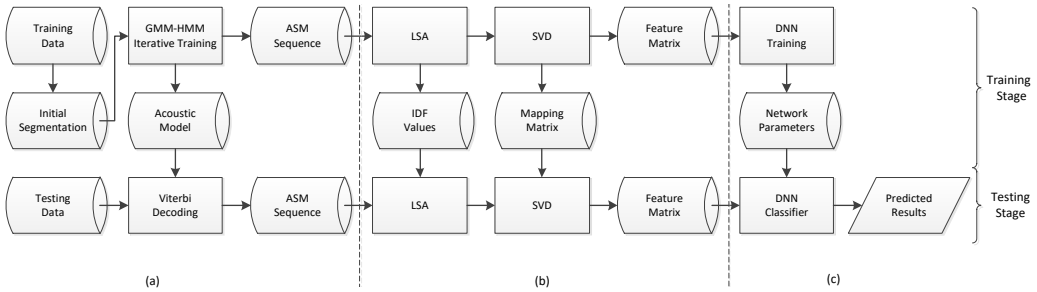

Figure 1: *Framework of our system based on ASM. (a)Generation of ASM sequences. (b)Feature vectors extraction by LSA. (c)Backend classifier*

Section 3, we discuss the experimental results and analysis. Finally, the summary of our work and the conclusion is proposed in Section 4.

2. The Proposed Architecture

This paper proposes a novel model for SER based on ASM. The framework is illustrated in Figure 1. By using ASM the speech with different emotions is transcRibed to ASM sequences. And the acoustic model which is generated in the iteration is used to get the ASM sequences of testing data. After decoding each acoustic recording is represented by a sequence of ASM units. Each unit can be regarded as a term in the text document. In information retrieval field, the relationship between document and term can be processed by topic models, such as LSA. The sequences of the speech are mapped to a term-document matrix based on LSA and each column is a fixed-length vector of a speech. After extracting the vectors in training data and testing data, these vectors are fed to DNN for classification.

2.1. Acoustic Segment Model

The purpose of using acoustic segment model is to convert each acoustic recording into a sequence of basic acoustic units, which is similar to the sentences consist of words. A typical ASM process involves two stages: initialization and model training. To get finer boundaries between the changes of acoustic features, we use the GMM-HMM-based method to explore the boundaries and divide the speech. The hidden states are the corpus to transcribe the audios to a sequence of ASM units.

2.1.1. Initialization

The core of ASM is the effectiveness of initial segmentation. In [19], just the simple even segmentation approach is used with an unsatisfactory result on SER. Because of the success in automatic speech recognition based on the GMM-HMMs, the method can model the segment of the speech well. As a reason, we use GMM-HMMs to segment the speech in the initialization stage.

First of all, The GMM-HMMs is used to model the acoustic speech. Just like modeling in automatic speech recognition, the HMM is a left-to-right topology. However, in the speech the similar emotion units may last for multi frames. So we increase a swivel structure in the topology in order to use the same hidden state to represent the similar frames. Suppose the dataset contains E emotions and there are N hidden states in a GMM-HMM. The Baum-Welch estimation[25] update the parameters of GMM-HMMs. After decoding, each acoustic recording can be represented by a sequence of hidden states and each hidden state is corresponding to a segment of the speech. Therefore, the $I = E \times N$ hidden states are the corpus to generate the initial ASM units.

2.1.2. Model Training

After the first initialization stage, each acoustic recording is represented by a sequence of ASM units. In the model training stage, each ASM unit is modeled by the GMM-HMM with a left-to-right HMM topology. As the first stage, the parameters of the model are updated by the Baum-Welch estimation. Then the training data is decoded to new sequences of ASM units by the Viterbi algorithm. The new sequences are applied as the new ASM units of the speech to train the parameters of GMM-HMMs in the next iteration. Repeat the above process until the ASM sequences of training data converge.

2.2. Latent Semantic Analysis

After the ASM process, each acoustic recording is transcribed as sequences of ASM units. Inspired by the success of latent semantic analysis(LSA) in information retrieval field, in our work, the ASM units are regarded as the terms in text document. LSA can produce a term-document matrix corresponding to the ASM units and acoustic recordings. Each column of the matrix is corresponding to the ASM sequences transcription of a recording while each row is corresponding to an ASM unit or binary of two ASM units.Therefore, if there are I terms of unigram, the dimension of vectors is $D = I \times (I + 1)$.

Like the processing of term-document matrix in information retrieval field, each element of the matrix in this work is generated by term frequency(TF) and inverse document frequency(IDF)[26]. The TF of a term is defined as the times of the term appears in the current document. And the IDF is the reciprocal of proportion of documents with this term in all documents. The pivotal segments that may affect emotions have more weight compared to the general segments have lower weight by using TF-IDF. The element in the i-th row and the j-th column which is the i-th term of the j-th acoustic recording of training data is defined by the following formulas:

$$TF_{i,j} = \frac{n_{i,j}}{\sum_{d=1}^{D} n_{d,j}} \tag{1}$$

$$IDF_i = log(\frac{M}{M(i)+1}) \tag{2}$$

where $n_{i,j}$ is the number of the i-th term in the ASM sequence of the j-th recording, M is the number of the training speech and the $M(i)$ is the number of documents in which the i-th term

appears. And the weight is the value of product of TF and IDF. The element of the matrix W is given by:

$$w_{i,j} = TF_{i,j} \times IDF_i \tag{3}$$

As the reason of using bigram, the term-document matrix W with demention of $D \times M$ is sparse. Then we use the SVD[27] to reduce the dimension of the matrix W as the formula:

$$W = U\Sigma V^T \tag{4}$$

The matrix W is decomposed into the product of three matrices: the left singular value $D \times D$ matrix U, the diagonal $D \times M$ matrix Σ which the diagonal elements are singular values arranged from large to small, and the right singular value $M \times M$ matrix V. Select the first k singular values of the Diagonal matrix and the first k rows of the left singular value matrix to form a reduced dimension mapping matrix U_k. Then the matrix multiply the diagonal matrix to the new matrix W_k with a low dimension. The matrix W_k is used to extract the feature vectors of training data. The value of k is designed by the percentage of sum of squares of singular values. In testing stage, the IDF and U_k trained before are used to generate the matirx W_k^{test}.

2.3. DNN Classifier

In this paper, DNN is used to classify the speech with different emotions. Due to the easily identifiable feature vectors for emotion classification extracted by ASM method, DNN is a simple three-tier structure.

3. Experiments

3.1. Database and Feature Extraction

The IEMOCAP database[28], widely used on the SER task, is adopted to validate our systems. There are five sessions in the IEMOCAP database and each session has speech utterances with emotion label of dialogs between two actors. Followed the evaluation protocol of [9, 29], we select the improvised data with four emotion categories i.e, happy, sad, angry and neutral. In this work, the audio recordings are processed by a sequence of overlapping Hamming windows with a 40-ms window size and a 20-ms window shift to extract 60-dimensional MFCC features. We design a five-fold cross validation. In each fold, the data from four sessions are training dataset, and the remaining session is split to two parts: the speech utterances of one actor are for validation and the remaining are the tesing dataset.

3.2. Experiments Results and Analysis

In this subsection, we explore different parameter settings of our system. The number of ASM units affects the precision of model partition, and the size of feature dimension reduction affects the representation and sparsity of feature vectors. These are the key to our system and we discuss the two parameters as follows.

3.2.1. Number of ASM units

The number of ASM units is the key to effective segmentation of the speech. Too few units are not sufficient to distinguish the boundaries of emotional change in a recording and too many units make the increase of computational complexity and the possibility of overfitting. Table 1 lists the result of different ASM units from 8 to 20. In this experiment the singular values of the first 80% of SVD are retained for the new matrix of reduced dimension. It is clear that 12 ASM units are most suitable for the segmentation on the SER task.

Table 1: *The accuracy comparisons of ASM units in different number.*

ASM units	Weighted Accuracy	Uweighted Accuracy
8	65.9%	61.7%
12	**72.6%**	**69.3%**
16	71.3%	67.2%
20	69.8%	64.1%

3.2.2. Dimension Reduction in SVD

In our work, unigram and bigram counts are used to generate terms in the speech. As a reason, the term-doucument matrix is sparse. In order to reduce the sparsity of matrix, the matrix only retains the largest singular values after SVD and the dimension of new matrix is decided by the percentage of sum of squares of singular values. Different percentage determines how much information the new matrix retains. In the experiment, the number of ASM units is 12. Therefore, the original diagonal matrix dimension is 156×156 in theory. However, the dimension is 152×152 because some terms of bigram are not existing. Table 2 displays the results with different percentage. As a result, the accuracy of SER is the highest while the percentage is 80%.

Table 2: *The accuracy comparisons of different reduced dimensions of SVD.*

Percentage	Weighted Accuracy	Uweighted Accuracy
70%	70.2%	66.8%
80%	**72.6%**	**69.3%**
90%	69.3%	65.4%
100%	71.9%	67.6%

3.2.3. Overall Comparision

The publish state-of-the-art results in [9] using the IEMOCAP database are showed in Table 3. Compared with the attention based FCN model, the accuracy of our system is improved by 2.2% and 5.4% on WA and UA absolutely. In order to verify the recognition effect of the two systems on different emotions, we make a fusion of the two systems. The results show that the two models are complementary for the recognition of different emotional speech.

Table 3: *The accuracy comparisons between ASM-DNN and the other systems.*

System	Weighted Accuracy	Uweighted Accuracy
FCN+Attention Model[9]	70.4%	63.9%
Our ASM-DNN	72.6%	69.3%
Fusion Model	**73.9%**	**70.8**%

We list the accurancy of different emotions by these systems in Table 4. As showed in Table 4, our system has a higher accurancy for 'neutral' and 'happy' speech contrast with 'sad' and 'anger' speech are easier to be classified by the model in [9]. Therefor, the fusion model from the two models outperforms the method of single model.

Table 4: *The accuracy of different emotions*

System	anger	neutral	happy	sad
Model in [9]	**63.1%**	74.5%	11.5%	**86.5%**
Our ASM+DNN	58.5%	**84.6%**	**18.3%**	76.8%
Fusion Model	61.7%	84.3%	16.9%	82.1%

In order to understand more intuitively that the speech with different emotions can be modeled through the ASM sequences, we extract the feature vectors after LSA for visualization compared with the features which are fed to the attention based FCN[9] from the spectrogram directly in Figure 2. It shows different emotional recordings are separated by ASM without supervision. Therefore, a simple classifier can also get competitive accuracy on the SER task.

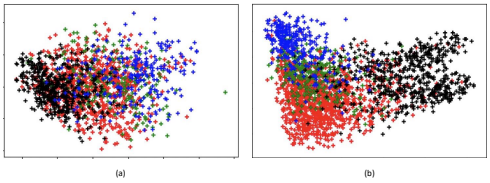

Figure 2: *Visualization results of feature vectors of the whole dataset. Blue, green, red, black is corresponding to 'angry', 'happy', 'neutral' and 'sad' respectively. (a) The feature vectors in [9]. (b) The feature vectors in our system.*

3.2.4. Results Analysis

Our system has a better accuracy of neutral and happy speech. Figure 3 and Figure 4 shows an example that the model in [9] confuses between the two emotions speech while our system distinguishes the two different emotions of the speech. For a more intuitive analysis, we show the decoding result sequences of the ASM units for the two example recordings and each recording is divided into a series of segments by the ASM sequences. The ASM units are named from S0 to S11. Clearly, the similar parts in spectrograms are decoded to the same ASM units, such as S4. And the difference of other ASM units like S7 in happy recording and S5 in neutral recording is the key to distinguishing the two emotional categories. In this way, our model segments the input audios in detail, thereby extracting more discriminative emotional acoustic features and outperforms the attention based FCN approach on the SER task.

The confusion matrix of emotions is occured in Figure 5. The speech in the category of 'happy' is hard to be recognised correctly while the accuracies of speeches contain 'neutral' and 'sad' are two highest. It is consistent with the conclusion of [30].

4. Conclusions

We demonstrated the acoustic segment model with the DNN classification achieves an outstanding performance on the SER task. Through the HMM-GMMs modeling the segmentation for the speech is very effective to distinguish different emotional units. In the backend, a simple deep neural network classifier is combined to divide the speech with different emotions. The traditional methods and deep learning are combined well on this task and the system outstrips the previous state-of-the-art accuracy on the IEMOCAP dataset. Besides, it's interesting the two systems are complementary in recognizing different emotional speeches and are worth further analysis and combination.

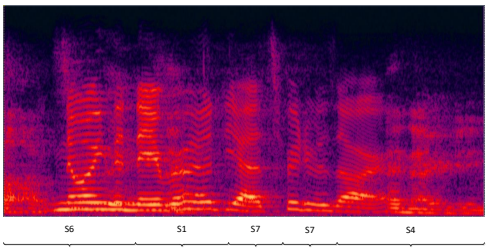

Figure 3: *The spectrogram and ASM sequences of an example recording of happy speech. This example was missclassified by attention based FCN model[9] as the neutral speech but correctly classified by our system.*

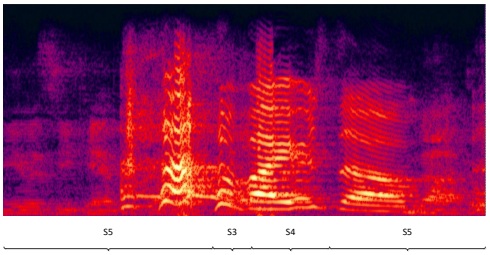

Figure 4: *The spectrogram and ASM sequences of an example recording of neutral speech. This example was missclassified by attention based FCN model[9] as the happy speech but correctly classified by our system.*

5. Acknowledgements

This work was supported in part by the National Key RD Program of China under contract No. 2017YFB1002202, the National Natural Science Foundation of China under Grants No. 61671422 and U1613211, the Key Science and Technology Project of Anhui Province under Grant No. 17030901005, and Huawei Noah's Ark Lab.

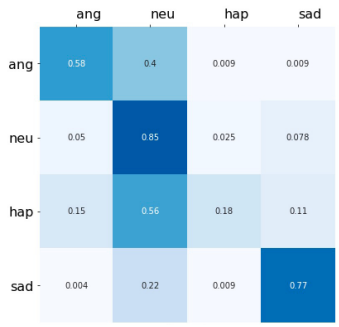

Figure 5: *The confusion matrix of different emotions.*

6. References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2002.

[2] F. Burkhardt, J. Ajmera, R. Englert, J. Stegmann, and W. Burleson, "Detecting anger in automated voice portal dialogs," in *Ninth International Conference on Spoken Language Processing*, 2006.

[3] C. Vinola and K. Vimaladevi, "A survey on human emotion recognition approaches, databases and applications," *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, vol. 14, no. 2, pp. 24–44, 2015.

[4] M. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[5] A. Hannun, C. Case, J. Casper, B. Catanzaro, and G. Diamos, "Deep speech: Scaling up end-to-end speech recognition," *Computer Science*, 2014.

[6] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2013.

[7] S. H. Bae, I. Choi, and N. S. Kim, "Acoustic scene classification using parallel combination of lstm and cnn," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, 2016, pp. 11–15.

[8] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Fifteenth annual conference of the international speech communication association*, 2014.

[9] Y. Zhang, J. Du, Z. Wang, J. Zhang, and Y. Tu, "Attention based fully convolutional network for speech emotion recognition," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 1771–1775.

[10] J. Wang, M. Xue, R. Culhane, E. Diao, J. Ding, and V. Tarokh, "Speech emotion recognition with dual-sequence lstm architecture," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6474–6478.

[11] E. Guizzo, T. Weyde, and J. B. Leveson, "Multi-time-scale convolution for emotion recognition from speech audio signals," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6489–6493.

[12] L. Pepino, P. Riera, L. Ferrer, and A. Gravano, "Fusion approaches for emotion recognition from speech using acoustic and text-based features," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6484–6488.

[13] D. Priyasad, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Attention driven fusion for multi-modal emotion recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3227–3231.

[14] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data," *Information Fusion*, vol. 49, pp. 69–78, 2019.

[15] C.-H. Lee, F. K. Soong, and B.-H. Juang, "A segment model based approach to speech recognition," in *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1988, pp. 501–541.

[16] H. Li, B. Ma, and C.-H. Lee, "A vector space modeling approach to spoken language identification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 271–284, 2006.

[17] M. Riley, E. Heinen, and J. Ghosh, "A text retrieval approach to content-based audio retrieval," in *Int. Symp. on Music Information Retrieval (ISMIR)*, 2008, pp. 295–300.

[18] X. Bai, J. Du, Z.-R. Wang, and C.-H. Lee, "A hybrid approach to acoustic scene classification based on universal acoustic models." in *INTERSPEECH*, 2019, pp. 3619–3623.

[19] H.-y. Lee, T.-y. Hu, H. Jing, Y.-F. Chang, Y. Tsao, Y.-C. Kao, and T.-L. Pao, "Ensemble of machine learning and acoustic segment model techniques for speech emotion and autism spectrum disorders recognition." in *INTERSPEECH*, 2013, pp. 215–219.

[20] T. Svendsen and F. Soong, "On the automatic segmentation of speech signals," in *ICASSP'87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 12. IEEE, 1987, pp. 77–80.

[21] T. J. Hazen, M.-H. Siu, H. Gish, S. Lowe, and A. Chan, "Topic modeling for spoken documents using only phonetic information," in *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2011, pp. 395–400.

[22] C.-T. Chung, C.-a. Chan, and L.-s. Lee, "Unsupervised discovery of linguistic structure including two-level acoustic patterns using three cascaded stages of iterative optimization," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8081–8085.

[23] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[24] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse processes*, vol. 25, no. 2-3, pp. 259–284, 1998.

[25] D. Elworthy, "Does baum-welch re-estimation help taggers?" *arXiv preprint cmp-lg/9410012*, 1994.

[26] D. Hull, "Improving text retrieval for the routing problem using latent semantic indexing," in *SIGIR'94*. Springer, 1994, pp. 282–291.

[27] M. E. Wall, A. Rechtsteiner, and L. M. Rocha, "Singular value decomposition and principal component analysis," in *A practical approach to microarray data analysis*. Springer, 2003, pp. 91–109.

[28] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[29] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms." in *Interspeech*, 2017, pp. 1089–1093.

[30] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1d & 2d cnn lstm networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, 2019.