SPEECH EMOTION RECOGNITION WITH I-VECTOR FEATURE AND RNN MODEL

Teng Zhang, Ji Wu

Multimedia Signal and Intelligent Information Processing Lab
Department of Electronic Engineering, Tsinghua University, Beijing, P.R.China

ABSTRACT

Machine-based emotion recognition from speech has emerged as an important research area in recent years. However, most studies have been done on artificial data. The difficulty of the recognition task increases when we facing natural speech data such as real-world conversations from call centre. Along with that difficulty, there are some new properties which may be useful to the real-world recognition tasks. In this paper, we focus on the recognition task on real-world conversations. Traditional prosodic acoustic features and the novel i-vector features are introduced and compared to represent the speech signal more abstractly. We also propose a Recurrent Neural Network approach to map the features to emotion labels. With only prosodic acoustic features and SVM multi-clasifier, we obtain a f-measure of 38.3%. By adding the i-vector features and the RNN model, we achieve a better result of 48.9%.

Index Terms— Emotion recognition,Speech analysis, Recurrent neural networks

1. INTRODUCTION

Speech signal contains a wealth of information about text, speaker, language, emotion and so on. Significant improvements have been achieved in speech recognition,speech retrieval and speaker identification. Meanwhile, emotion recognition from speech remains a challenge task due to the semantic gap between speech signal and people's cognition. People express their emotions by the manner they speak the words, which occupies only a small part of the entire speech. A straightforward idea to detect emotion from speech is to take out the text, spectrum, and other information from the speech, which is often described as a process of feature extraction.

Choosing suitable features is a crucial procedure for speech processing systems. Different features convey different messages in highly overlapped manner, which makes it difficult to determine which feature should be chosen to represent emotion. I.Iliev and Scordilis[1] have discussed the use of glottis excitation features in speech emotion recognition tasks. Vocal tract features such as MFCCs and PLPCs are more popular in emotion recognition tasks. Bitouk[2] uses MFCC features from consonant, stressed and unstressed vowels to realize the emotion classification on English LDC and Emo-DB databases. Degaonkar[3] introduces the wavelet packet coefficient to express the emotion information in speech. Prosody features are often associated with larger speech units such as syllables, words, phrases and even sentences. Cowie[4] has showed that pitch, loudness, speech rate, and spectrum are the most effective features for emotion recognition. In this paper, we use a combination of traditional prosody features and vocal tract features based on i-vector to represent the emotion information in speech.

When the effective features have been extracted, a classifier can be used to map the features to emotion labels. Zhou[5] uses the gaussian mixture models to model the articulatory features and obtain the recognition result. Kamaruddin[6] proposes a four-states Hidden Markov Model to take care of the effects of speaking rate and variation of F0. Fernandez[7] proves that an artificial neural network can be useful in emotion recognition tasks. The choice of classification method mainly depends on the data volume. Common emotional speech databases only contain hundreds of speech clips, which makes it difficult to use complex nonlinear classification models such as deep neural networks. In this paper, we acquire thousands of natural speech clips from a call centre. In consideration of the liberal quantity and temporal expression of the data, we propose a recurrent neural network approach to implement the recognition task.

The rest of this paper is organized as follows. In section 2, we state the data set used in our experiment and its acquisition. Section 3 discusses the prosody features and vocal tract features we select to represent the emotion information in speech. Then a recurrent neural network approach is introduced to implement recognition tasks in section 4. Section 5 conducts some experiments and evaluates the performance of the proposed system. At last, we conclude this paper and present our future work in Section 6.

2. DATA ACQUISITION

The data set used in this study is collected from a practical scenario and manually labeled for our study. In a call-centre that handled customer inquiries for several electricity companies, customers call and speak to a customer service representative to ask help for some problems or acquire informa-

tion and service. If their problems are resolved perfectly, they may be happy, otherwise they may be dissatisfied, sometimes even feel angry. The attitude of a customer service representative can also influence the customers' emotion.

For this database, a total of 941 conversations between customers and customer service representatives are selected. The average length of the conversations is 136.12 s, the longest conversation lasts for 976.22 s, the shortest one is 57.9 s. Because of the obviously different acoustic properties between customers and customer service representatives, we firstly remove the customer service representatives' component from the conversations. Then each conversation file was manually segmented into sentence-level utterances. The maximum utterance duration is 53.15 s, the minimum is 0.26 s, and the mean value of duration is 4.34 s. Each utterance is manually labeled into three types of emotion, $positive$ which means customers are calm or their problems have been resolved perfectly, $middle$ which means they are a little bit impatient or dissatisfied for the service, and $negative$ which means they are angry or uncontrollable. Finally we get a total of 16073 utterances in the database where only 1.9% of them are negative samples, 19.7% of them are middle samples, and the rest 78.4% are positive samples.

3. ACOUSTIC FEATURES

Previous studies have proved that prosody and spectrum features perform well in speech emotion recognition tasks. In this section, we first make use of some common prosody and spectrum features that can be easily extracted with the $OpenSmileToolkit$[10], then the i-vector based spectrum features are extracted to improve the effectiveness of acoustic features.

3.1. *Fundamental features*

Based on previous studies on acoustic features associated with emotion, we select features from four prosodic groups: *the fundamental frequency*, *energy*, *speaking rate*, and *spectrum*. The fundamental frequency and energy are extracted and linked as curves. we get *mean*, *standard deviation*, *median*, *maximum*, *minimum*, *range*, *quartiles*, and *distances* of the curves as the initial features. For the speaking rate, we get the the proportion of voiced clips in the whole speech. The mean and standard deviation of 13 mel-frequency cepstral coefficients are used to represent the spectrum information.

In addition, we found in our research work that the going up and down of the fundamental frequency were also helpful to emotion recognition.

3.2. *I-vector*

The i-vector technique[11] is developed based on Joint Factor Analysis[12]. It is widely used in speaker recognition and language identification tasks. Owing to its good performance on information separation, it is also used in some other speech understanding tasks in recent years. Huang[13] has used the i-vector technique to detect acoustic event in multimedia event detection task conducted annually by National Institute of Standards and Technologys(NIST). Xia[14] has tried to apply it to some simple emotion recognition tasks.

Giving thousands of sentence-level utterances, we can have many choices among the probabilistic models to describe the speech feature space, such as Hidden markov model(HMM), Gaussian mixture model(GMM) and so on. Douglas[15] showed that when our task is text-independent, GMM may be better. So we can roughly use a GMM with k components to describe the speech space, which is called as the universal background model(UBM). Using the UBM, we can get a super-vector m by concatenating mean vectors of these gaussian densities for all utterances. This part can be described as follows.

1. Initialize the GMM mean and covariance matrix m_k and Σ_k
2. E-step: Compute the generating probability of each utterance x_i from each GMM component c using Formula (1)

$$\gamma(i,c) = \frac{\pi_c N(x_i|\mu_c, \Sigma_c)}{\sum_{j=1}^{k} \pi_j N(x_i|\mu_j, \Sigma_j)} \tag{1}$$

3. M-step: Update GMM mean and covariance matrix m_k and Σ_k using Formula (2)

$$\begin{aligned} N_c &= \sum_{i=1}^{N} \gamma(i,c) \\ \mu_c &= \frac{1}{N_c} \sum_{i=1}^{N} \gamma(i,c) x_i \\ \Sigma_c &= \frac{1}{N_c} \sum_{i=1}^{N} \gamma(i,c)(x_i - \mu_c)(x_i - \mu_c)^T \end{aligned} \tag{2}$$

4. Repeat 2 and 3 until m_k and Σ_k no longer changes

Then for each utterance, there is a posterior super-vector μ that is a little deviated from m. For the formula $\mu - m = Vy$, we can use an expectation-maximization(EM) procedure to estimate the eigenmatrix V and the i-vector y.

From the procedure above, we can see that the i-vector technique estimate the difference between the real data and the "average data", which means a somehow smoothness on speaker, language, and other irrelevant information in speech. The experiments in Section 5 prove that i-vector can obtain better emotion recognition results than the fundamental features. Furthermore, i-vector features are closer to the spectrum description of the speech in some ways, which may be

complementary with the fundamental prosodic features. So we also combine these two kinds of features and get some improvement in our experiments.

4. RECOGNITION MODELS

As described in Section 2, the database used for this study consists of 16073 speech utterances segmented from 941 continuous conversations. The utterances spoken by one person within a certain amount of time must have some relation in both signal level and emotion labels, which has the flavor of markov property. Instead of simply recognize the emotion label of each utterance, we extend the problem to recognize a emotion sequence S of utterances O in the same conversation, as Formula (3) describes.

$$\hat{S} = \arg\max_S p(S|O) = \arg\max_S p(O|S)p(S) \tag{3}$$

Let's make a distinction between the two components of Formula (3), the previous part $p(O|S)$ is called the classification model which will be described in detail in this section and the latter part $p(S)$ is called the emotion model which is only related to the emotion space.

Assume that the observation of each utterance o_t determined by the current emotion label s_t, we can rewrite $p(O|S)$ as Formula (4).

$$p(O|S) = \prod_t p(o_t|s_t) \propto \prod_t \frac{p(s_t|o_t)}{p(s_t)} \tag{4}$$

where $p(s_t|o_t)$ is a simple classification problem and $p(s_t)$ is the prior probability which can be counted using the database.

Support vector machine[16] is widely used in classification and recognition task. So we take it as the basic classifier. As we have discussed before, the temporal relations in signals should not be ignored. A straightforward method is to compute the difference between neighbouring utterances as additional features participating in the classification process. We want to use a more powerful model, the recurrent neural network(RNN), which will be proved more effective in Section 5.

Recurrent neural network(RNN)[18] is an artificial neural network model that are well-suited for pattern classification tasks whose inputs and outputs are sequences. The standard RNN is a nonlinear dynamical system that maps sequences to sequences, as shown in Figure 1.

The parameters of RNN consist of three transition matrices and two bias vectors, W_{hv} which represents the connection matrix between the input features and the hidden states, W_{hh} which represents the connection matrix between the hidden states and the former ones, W_{hz} which represents the connection matrix between the hidden states and the output labels, b_h which represents the offset in the hidden states

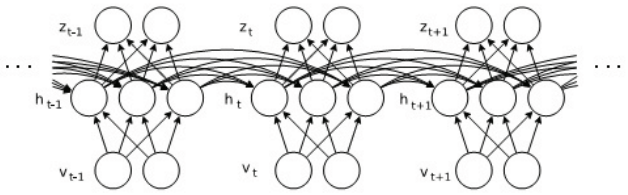

Fig. 1: Recurrent neural network

and b_z which represents the offset in the output labels. Given an input feature $V = (v_1, v_2, ..., v_T)$ and all the parameters, the RNN computes the output emotion labels as follows.

Algorithm 1 The forward procedure

1: **for** each $t \in [1, T]$ **do**
2: $u_t = W_{hv}v_t + W_{hh}h_{t-1} + b_h$
3: $h_t = e(u_t)$
4: $o_t = W_{hz}h_t + b_z$
5: $z_t = g(o_t)$
6: **end for**

$e(\cdot)$ is the hidden non-linear function. There are mainly three types of non-linear functions, the sigmoid function, the tanh function and the maxout function. $g(\cdot)$ is the output non-linear function. Assume that o_t is a k-dimension vector, in this paper, we use a softmax regression model[19] to compute the probability of emotion labels as Formula (5). Then a cost function based on the logistic regression can be computed as Formula (6).

$$p(z_t = c|v_t) = \frac{e^{o_t(c)}}{\sum_{i=1}^k e^{o_t(i)}} \tag{5}$$

$$J = -\frac{1}{T}[\sum_{i=1}^{T}\sum_{j=1}^{k} 1\{z_t = j\} log p(z_t = j|v_t)] \tag{6}$$

Giving the database for training, we can use a stochastic gradient descent method[20] to estimate the RNN parameters. The derivatives of the RNN are easily got with the back-propagation through time algorithm[21].

5. EXPERIMENTS

we conduct some experiments to verify the performance of our emotion recognition system. First we survey the performance of the acoustic features we have defined in Section 3, then a RNN is trained to get the optimal emotion sequences.

The database used for this study consists of 16073 utterances from 941 conversations. We divide it into three parts, the training part, the developing part and the testing part.

Algorithm 2 BPTT algorithm

for each $t \in [T, 1]$ **do**
  $do_t = g^{'}(o_t)dJ/dz_t$
  $db_z = db_z + do_t$
  $dW_{hz} = dW_{hz} + do_t h_t^T$
  $dh_t = dh_t + W_{hz}do_t$
  $dz_t = e^{'}(z_t)dh_t$
  $dW_{hv} = dW_{hv} + dz_t v_t^T$
  $db_h = db_h + dz_t$
  $dW_{hh} = dW_{hh} + dz_t h_{t-1}^T$
  $dh_{t-1} = W_{hh}dz_t$
end for

Table 1: **Performance(F-measure) of different features**

Feature group	Neg(%)	Mid(%)	Pos(%)	Mean(%)
Fundamental	7.6	36.4	70.9	38.3
I-vector	14.2	36.2	77.2	42.5
Combination	**15.0**	**37.1**	**77.7**	**43.3**

5.1. *Experiments on different features*

Two kinds of acoustic features were used in our study, the fundamental features we had defined in Table 1 and the i-vector. We used the $OpenSmileToolkit$ to extract fundamental features and then trained a multi-class SVM classifier to recognize the emotion of speech utterances. This was our baseline system.

For the i-vector, a UBM was trained first using part of the training data, which based on a GMM with 64 mixtures. The dimension of i-vector, which is equivalent to the rank of eigenmatrix V, was chosen to be 64 because our experiments showed that lager dimension could not provide any performance improvement.

We also experimented a combination of these two kinds of features, which achieved the best performance in this section. The performance of different features could be seen in Table 1.

The combination features provided the best performance for all three emotions, and we would use it in the next experiments.

5.2. *Experiments using Recurrent neural network*

In this part, we used a recurrent neural network to replace the SVM classifier used in Section 5.1. For a recurrent neural network, in addition to the parameters described in Section 4, there were still a few important structure-related parameters that should be confirmed during the experiments. First was the number of hidden notes n, if n was too small, the representation capability of the network would be limited, however, if n was too large, it became difficult to converge to a stable result. Next was the type of hidden non-linear function, we should select a proper function for our emotion recognition tasks. Finally was the iteration times, we should get a balance of performance and over-fitting.

Table 2: **Performance(F-measure) of Recurrent Neural Network**

Num of nodes	Neg(%)	Mid(%)	Pos(%)	Mean(%)
3	24.9	**37.5**	**84.2**	**48.9**
5	**25.3**	36.8	83.0	48.4
10	22.2	34.5	83.2	46.6
Best SVM	15.0	37.1	77.7	43.3

We used training set to train the parameters in Section 4, and used the developing set to determine additional parameters above. As a result, we used the tanh hidden non-linear function, $n = 3$ and stop our iteration process when the performance in the developing set was the best. The results were listed in Table 2.

6. CONCLUSION

In this paper, we explored the performance of two different kinds of acoustic features in the emotion recognition task. As seen in paper, the i-vector features gave a performance improvement from 38.3% to 42.5%, and a simple combination of them obtained a better performance of 43.3%. This result proved that i-vector features are effective in representing the emotion information in speech.

This paper also proposed a Recurrent neural network(RNN) approach to make use of the temporal properties of speech signals and emotion labels. RNN improved the performance from 43.3% to 48.9%.

However, the emotion recognition task for speech is still far from solved. What features can best represent the emotion information? What kind of model can best describe the emotion space? Both questions are up in the air. Our future work will go to the bottom of the signal-level or go to the top of the emotion space to explore the interesting cognitive problems.

7. REFERENCES

[1] Iliev A I, Scordilis M S. Spoken emotion recognition using glottal symmetry[J]. EURASIP Journal on Advances in Signal Processing, 2011, 2011: 2.

[2] Bitouk D, Verma R, Nenkova A. Class-level spectral features for emotion recognition[J]. Speech Communication, 2010, 52(7): 613-625.

[3] Degaonkar V N, Apte S D. Emotion modeling from speech signal based on wavelet packet transform[J]. International Journal of Speech Technology, 2013, 16(1): 1-5.

[4] Cowie R, Douglas-Cowie E, Tsapatsoulis N, et al. Emotion recognition in human-computer interaction[J]. Signal Processing Magazine, IEEE, 2001, 18(1): 32-80.

[5] Zhou Y, Sun Y, Yang L, et al. Applying articulatory features to speech emotion recognition[C]//Research Challenges in Computer Science, 2009. ICRCCS'09. International Conference on. IEEE, 2009: 73-76.

[6] Kamaruddin N, Wahab A. Features extraction for speech emotion[J]. Journal of Computational Methods in Science and Engineering, 2009, 9: 1-12.

[7] Fernandez R, Picard R W. Modeling drivers speech under stress[J]. Speech Communication, 2003, 40(1): 145-159.

[8] Meng Q, Wu W. Artificial emotional model based on finite state machine[J]. Journal of Central South University of Technology, 2008, 15: 694-699.

[9] Koolagudi S G, Rao K S. Emotion recognition from speech: a review[J]. International Journal of Speech Technology, 2012, 15(2): 99-117.

[10] Guyon I, Elisseeff A. An introduction to variable and feature selection[J]. The Journal of Machine Learning Research, 2003, 3: 1157-1182.

[11] Dehak N, Kenny P, Dehak R, et al. Front-end factor analysis for speaker verification[J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2011, 19(4): 788-798.

[12] Kenny P. Joint factor analysis of speaker and session variability: Theory and algorithms[J]. CRIM, Montreal,(Report) CRIM-06/08-13, 2005.

[13] Huang Z, Cheng Y C, Li K, et al. A blind segmentation approach to acoustic event detection based on i-vector[C]//INTERSPEECH. 2013: 2282-2286.

[14] Xia R, Liu Y. Using i-Vector Space Model for Emotion Recognition[C]//INTERSPEECH. 2012.

[15] Reynolds D A, Quatieri T F, Dunn R B. Speaker verification using adapted Gaussian mixture models[J]. Digital signal processing, 2000, 10(1): 19-41.

[16] Cortes C, Vapnik V. Support-vector networks[J]. Machine learning, 1995, 20(3): 273-297.

[17] Gross J J. Emotion regulation in adulthood: Timing is everything[J]. Current directions in psychological science, 2001, 10(6): 214-219.

[18] Sutskever I. Training recurrent neural networks[D]. University of Toronto, 2013.

[19] Gold S, Rangarajan A. Softmax to softassign: Neural network algorithms for combinatorial optimization[J]. Journal of Artificial Neural Networks, 1996, 2(4): 381-399.

[20] Zhang T. Solving large scale linear prediction problems using stochastic gradient descent algorithms[C]//Proceedings of the twenty-first international conference on Machine learning. ACM, 2004: 116.

[21] Werbos P J. Backpropagation through time: what it does and how to do it[J]. Proceedings of the IEEE, 1990, 78(10): 1550-1560.

[22] Song F, Croft W B. A general language model for information retrieval[C]//Proceedings of the eighth international conference on Information and knowledge management. ACM, 1999: 316-321.

[23] Paul D B. An efficient A* stack decoder algorithm for continuous speech recognition with a stochastic language model[C]//Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on. IEEE, 1992, 1: 25-28.

[24] Burkhardt F, Paeschke A, Rolfes M, et al. A database of German emotional speech[C]//Interspeech. 2005, 5: 1517-1520.

[25] Kotti M, Patern F. Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema[J]. International journal of speech technology, 2012, 15(2): 131-150.

[26] Yang B, Lugger M. Emotion recognition from speech signals using new harmony features[J]. Signal Processing, 2010, 90(5): 1415-1423.

[27] Wang K, An N, Li L. Speech emotion recognition based on wavelet packet coefficient model[C]//Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on. IEEE, 2014: 478-482.