EMOTIONAL SPEAKER RECOGNITION BASED ON I-VECTOR THROUGH ATOM ALIGNED SPARSE REPRESENTATION

*Li Chen, Yingchun Yang**

Zhejiang University, College of Computer Science & Technology, Hangzhou, China
{stchenli,yyc}@zju.edu.cn

ABSTRACT

I-vector algorithm was previously adopted to improve the performance of ASR (Automatic Speaker Recognition) system which is degraded by emotion variability. The variability compensation technique is LDA (Linear Discriminant Analysis) which assumes the variability is speaker-independent. However, this assumption is not suitable for emotion variability because we discover that the pattern of emotion variability is speaker-dependent. Therefore, a novel emotion synthesis algorithm AASR (Atom Aligned Sparse Representation) is proposed to characterize this speaker-dependent pattern and compensate the emotion variability within i-vectors. The experiments conducted on MASC show that our algorithm, compared with the GMM-UBM algorithm and the conventional variability compensation algorithm LDA, both can enhance the speaker identification and verification performances.

Index Terms— Emotional Speaker Recognition, Speaker-Dependent Variability, Atom Aligned Sparse Representation

1. INTRODUCTION

People often speak with different emotional states in real-life yet the enrollment speeches of the speaker identification application are usually neutral, resulting in the emotion inconsistency between enrollment and test data. This inconsistency degrades the performance of ASR system, and ESR (Emotional Speaker Recognition) is proposed for alleviating this negative effect.

Many efforts have been taken to improve the performance of ESR system. Recently, the methods used to compensate the channel variability are applied to eliminate the effect induced by emotional state mismatch. EAP (Emotion Attribute Projection) [1] is developed on the basis of NAP (Nuisance Attribute Projection) to subtract the emotional attribute from each utterance. Also, i-vector [2] modeling the total variability (containing channel variability and speaker variability) is applied into ESR by replacing the channel variability matrix with the emotion variability matrix [3]. Both methods can enhance the performance of ESR compared with the algorithms without emotion compensation.

SRC (Sparse Representation Classifier) is also applied into speaker recognition system. The fixed-length representation of a given utterance can be considered as a sparse linear sum of the atoms of the overcomplete dictionary. Naseem introduced SRC into speaker recognition on a pure telephone database – TIMIT [4]. The GMM mean supervector is utilized to represent an utterance. Afterwards, SRC based on the low-dimensional representation – i-vector is proposed [5], and SRC is used to compute the similarity score of two i-vectors, based on the process of channel compensation.

In conventional i-vector algorithm, the common variability compensation strategy is LDA and WCCN (Within Class Covariance Normalization) based on the assumption that the pattern of variability is speaker-independent. However, the assumption is not suitable for the emotion mismatch problem because we discover that the pattern of emotion variability is dependent on the speakers. Thus, we propose a novel compensation algorithm named AASR (Atom Aligned Sparse Representation) to model the speaker-dependent relationship between the total variability space and the pure speaker space. Meanwhile, the emotion synthesis algorithm based on AASR is used to substitute the conventional variability compensation method LDA. Our algorithm achieves a promising result on the emotional corpus MASC.

The remainder of this paper is organized as follows. The traditional speaker recognition system based on i-vector is introduced in Section 2. Section 3 illustrates the Sparse Representation algorithm. Section 4 explains the speaker-dependent property of emotion variability and presents the novel AASR algorithm. The experimental results on MASC are reported in Section 5. Finally, the conclusion is drawn in Section 6.

2. I-VECTOR ALGORITHM

The conventional i-vector algorithm, used to solve the channel mismatch problem, usually takes three steps:

(1) i-vector extraction. The i-vector extraction can be seen as a probabilistic compression process which reduces the di-

*Thanks to 973 Program 2013CB329504, the Fundamental Research Funds for the Central Universities 2013 and National Natural Science Foundation of China (NSFC60970080) for funding.

mensionality of the GMM mean supervector. It models the speaker-dependent and channel-dependent GMM supervector $M_{(s,h)}$ as the sum of the speaker-independent mean supervector m and total variability vector,

$$M_{(s,h)} = m + Tw_{(s,h)}, \tag{1}$$

where m is the UBM (Universal Background Model) mean supervector. T and $w_{s,h}$ represent the total variability matrix and the speaker-dependent and channel-dependent i-vector respectively.

(2) variability compensation. Since the i-vector extraction algorithm does not separate the speaker variability and the channel variability, channel compensation method must be applied before computing the similarity score of two i-vectors. The traditional compensation method is the combination of LDA and WCCN which remove the channel attribute from i-vector.

The LDA algorithm maps the original space to the reduced optimal space by minimizing the intra-speaker variability and maximizing the inter-speaker variability,

$$\arg\max_{u} \frac{u^t S_b u}{u^t S_w u}, \tag{2}$$

where u is the direction of LDA projection matrix A, and S_b and S_w are the inter-speaker covariance matrix and the intra-speaker covariance matrix respectively. LDA linearly transforms the original total variability space into the speaker variability space. LDA works well on the assumption that the transformation relationship between these two spaces is class-independent, i.e., speaker-independent in ASR system. Channel variability can be regarded as speaker-independent according to the analysis in [6].

The WCCN algorithm is to normalize the covariance of the dimension-reduced i-vector to minimize the error expectation of false alarm and false rejection.

(3) score computation. CDS (Cosine Distance Scoring) method is applied to compute the similarity between two i-vectors x_1 and x_2,

$$s(x_1, x_2) = \frac{(A^t x_1)^t W^{-1} (A^t x_2)}{\sqrt{(A^t x_1)^t W^{-1} (A^t x_1)} \sqrt{(A^t x_2)^t W^{-1} (A^t x_2)}}, \tag{3}$$

where A is the LDA projection matrix and W is the WCCN matrix.

The i-vector algorithm is also applied into ESR by replacing the channel compensation with the emotion compensation when LDA algorithm is implemented in [3]. However, the emotion variability is somewhat different from the channel variability because the pattern of emotion variability is speaker-dependent.

3. SPARSE REPRESENTATION ALGORITHM

SR (Sparse Representation) algorithm assumes that a signal can be represented by a sparse combination of some redundant bases (i.e. atoms) which constitute a dictionary. The typical SR algorithm usually takes two steps:

(1) Constructing the dictionary based on the development corpus. The dictionary can be an exemplar-based dictionary or a learned dictionary. An exemplar-based dictionary arranges the i-vectors of the speakers as its atoms. A learned dictionary can be trained in various ways. The learning algorithm, mentioned in [7], is the feature-sign search and Lagrange Dual algorithm. The dictionary D is trained by solving Equation(4),

$$\arg\min_{C,D} ||Y - DC||_2^2 + \gamma ||C||_1, \tag{4}$$

where Y is the set of training vectors, D is the resultant dictionary consisting of the basis vectors of the linear space spanned by Y, C is the set of sparse coefficient vectors corresponding to the set of Y, and γ is the weight which is used as the tradeoff between the regression error $||Y - DC||_2^2$ and the sparsity $||C||_1$ and is set to 5 according to our experiments.

(2) Attaining the sparse coefficient of the test samples. It is a key step of SR algorithm and is achieved by L1-norm technique,

$$\arg\min_{c} ||c||_1 \quad subject\,to \;\; ||y - Dc||_2^2 \leq \varepsilon, \tag{5}$$

where y is the test sample, c is the resultant sparse coefficient, D is the dictionary trained in step (1) and ε is the error bound.

SR can be used in various ways. In [5], the regression error $||y - D\delta(c)||_2^2$ is applied as the distance of the test sample and target speakers' atoms. $\delta(c)$ indicates that the coefficients, not corresponding to the target speaker, are set to zero. In our paper, SR is applied for model synthesis through aligned dictionaries.

4. ATOM ALIGNED SPARSE REPRESENTATION

4.1. Speaker-Dependent Emotion Variability

We draw the vowel triangle to observe the variation of formants under neutral and panic between two different male speakers as in Fig. 1, which illustrates that the amplitude and the direction of the variation are different between two speakers. Thus, the pattern of variation may vary among different speakers. This opinion is also supported by many other studies on emotion recognition [8]. The dependency is mainly caused by two factors. First, the channel variability is objective while the emotion variability is subjective, because the emotional strength and expression style vary from person to person. Second, the channel of an utterance is certain while the emotional state is not because each person has different emotional perception. Thus, we assume that the relationship of emotional i-vector y_e and neutral i-vector y_n is modeled as a speaker-dependent transformation function f_s,

$$y_e = f_s(y_n). \tag{6}$$

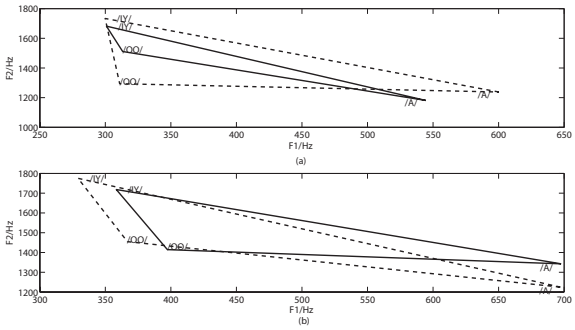

Fig. 1. Vowel triangle of two male speakers under neutral and panic. (a) and (b) represent two different male speakers. The solid and dot lines represent the vowel triangle under neutral and panic respectively. The phonemes corresponding to the 3 vertexes of the vowel triangle are $/A/$, $/OO/$, $/Y/$ respectively.

According to the analysis, LDA is not quite suitable for emotion compensation so we develop a novel emotion synthesis algorithm named AASR (Atom Aligned Sparse Representation). The AASR is inspired by SR algorithm used to recover and synthesize facial-expression mentioned in [9] and it can depict the speaker-dependent pattern of emotion variability.

4.2. Atom Aligned Sparse Representation

In our development corpus, each neutral utterance has an aligned emotional utterance, which means that the two utterances are from the same speaker and the contents of the pair are the same. Given M aligned neutral and emotional i-vector pairs $(y_{i,n}, y_{i,e}), i = 1, 2, \ldots, M$, the AAD (Atom Aligned Dictionary) D can be trained by Equation(7), which can be solved by feature-sign search algorithm and Lagrange Dual [7].

$$\arg\min_{C,D} ||Y - DC||_2^2 + \gamma||C||_1. \tag{7}$$

The variable Y is the set of combined neutral and emotion pairs: $Y = \left(y_{:,n}^T, y_{:,e}^T\right)^T$. $D = (D_n^T, D_e^T)^T$ is the AAD. The sharing coefficient C_i can represent both $y_{i,n}$ and $y_{i,e}$ based on the neutral dictionary D_n and emotional dictionary D_e respectively. The coefficient-sharing property indicates that the aligned i-vectors $(y_{i,n}, y_{i,e})$ share the same coordinates under the spaces spanned by D_n and D_e. Thus, corresponding atoms $D_{j,n}$ and $D_{j,e}$ representing the bases of these two spaces can be regarded as aligned. The schematic diagram of the process is shown in Fig. 2.

D_n is regarded as the speaker variability space because no emotion variability exists in D_n, and D_e represents the total variability space, containing the emotion variability and the

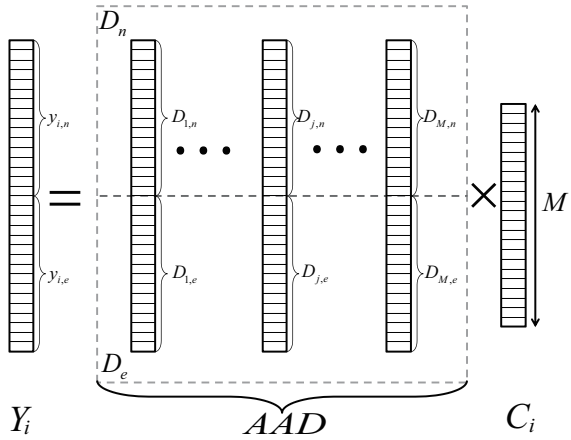

Fig. 2. Schematic diagram of AASR. Y_i is the ith training pair from Y and C_i is Y_i's corresponding sparse coefficient. The AAD is the resultant aligned dictionary where the atom pair ($D_{j,n}$,D_j, e) is aligned.

speaker variability. The corresponding bases $D_{j,n}$ and $D_{j,e}$ indicate a transformation law from a base vector of the pure speaker space to that of the total variability space. The latent transformation law f_j is

$$D_{j,e} = f_j(D_{j,n}). \tag{8}$$

f_j is different for each $D_{j,n}$, so f_j can model the speaker-dependent transformation relationship. If c_n is the sparse representation of a neutral i-vector y_n on the dictionary D_n by Equation (5), c_n is used to represent the speaker's emotional i-vector y_e based on the emotional dictionary D_e.

$$y_e = D_e * c_n. \tag{9}$$

This transformation strategy can also be interpreted as that the jth transformation law f_j corresponding to the nonzero value $c_{n,j}$ is used to guide the synthesis of emotional i-vector y_e because $c_{n,j}$ indicates y_n is related to the jth atom in D_n.

4.3. Application of AASR

Although the emotion variability is much more complex than the channel variability, there is one advantage that the number of emotional states is less than that of channels. Because the channels with different telephones or microphones are regarded as different and the number of channels are innumberable, only the channel removal method LDA can be used. In contrast, there are several theories about key emotional states which can code the emotions expressed in speech. We adopt the four key emotional states theory (anger, elation, panic and sadness) mentioned in [10]. Thus, the emotion synthesis method is used to substitute the emotion removal method LDA mentioned in step (2) of Section 2. The synthesis algorithm introduced in Section 4.2 is used to generate the emotional

Table 1. IRs by GMM-UBM, i-vector and AASR(%).

	GMM-UBM	i-vector	AASR
neutral	96.23	93.47	96.80
anger	31.50	48.83	50.03
elation	33.57	49.40	50.90
panic	35.00	42.77	47.37
sadness	61.43	64.93	66.97
average	51.55	59.88	62.41

i-vector from the speaker's neutral one. In evaluation stage, each test sample is scored not only on the neutral i-vector of the target speaker, but also on the synthesized four emotional i-vectors by the CDS score computation method. The final score is the maximum value of these scores.

5. EXPERIMENT

5.1. Corpus and Setup

We perform our experiments on an emotional speech database named MASC (Mandarin Affective Speech Corpus) [11], which includes five emotional states: neutral, anger, elation, panic and sadness. 68 native Mandarin speakers (23 females and 45 males) are asked to produce 20 utterances for three times under these five emotional states and 2 extra neutral paragraphs. Each neutral utterance of a speaker has a corresponding utterance with the same content under each emotional state. The utterance pairs thus can be constructed by the corresponding neutral and emotional utterances. In our experiments, the first 18 speakers are used as development corpus and the remaining 50 speakers as evaluation corpus. Each target speaker's model in the evaluation corpus is trained by using the neutral paragraphs and all the utterances are used as test samples.

In our experiments, 13-order MFCC (Mel-Frequency Cepstrum Coefficient) plus delta is extracted with 32ms frame length at a 16ms frame rate. The neutral paragraphs of the development corpus are utilized to train the 512-component UBM by Expectation Maximize (EM) algorithm.

The total variability matrix T is trained by all the utterances and paragraphs in the development corpus. The 300-dimension i-vector of each sample is extracted. For conventional i-vector, LDA strategy is applied to reduce the i-vector's dimensionality to 200. For AASR, the AAD is trained by using all the aligned pairs in the development corpus and the pair number under each emotional state is 1080.

5.2. Experimental Result

The IRs (Identification Rates) of the GMM-UBM algorithm, the i-vector algorithm with LDA and WCCN compensation technique, and our AASR algorithm are shown in Table. 1.

As shown in Table. 1, our algorithm can enhance the performance of ASR in every emotional state. The average IR increases by 10.86% and 2.53%, compared with that of GMM-UBM and conventional i-vector algorithm. It is also worth mentioning that the conventional i-vector algorithm would degrade the performance of ASR under neutral yet our algorithm does not. Because the neutral enrollment utterances and neutral test utterances are matched and the emotion compensation process is not expected, our emotion synthesis method won't change the original neutral model while the variability removal method LDA removes the emotion attribute from the neutral i-vector.

The DET (Detection Error Tradeoff) curve of the three algorithms are shown in Fig. 3, and the improvement of EER (Equal Error Rate) is similar to that of average IR.

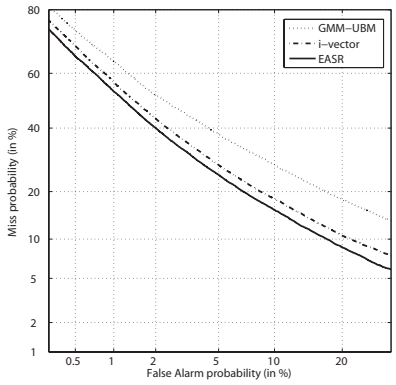

Fig. 3. DET plot by GMM-UBM, i-vector and AASR.

The EERs are 18.75%, 14.10% and 12.77% for GMM-UBM, i-vector and AASR algorithm respectively. Our algorithm also outperforms the two baseline algorithms. The EER decrease by 5.98% and 1.33%, compared with that of GMM-UBM and conventional i-vector.

6. CONCLUSION

Emotion variability has some common property with the channel variability, yet they have many distinct properties. This paper discovers that the channel variability may be speaker-independent yet the emotion variability is speaker-dependent. Thus, the common channel compensation technique LDA is not suitable for ESR system. We propose a novel speaker-dependent emotion compensation technique – AASR to synthesize the emotional i-vectors. In the future, we will employ more effective manifold tools to depict the concrete dependency and propose explicit emotion compensation technique, such as rule based compensation, to handle speaker-dependent variability.

7. REFERENCES

[1] H. Bao, M. Xu, and T.F. Zheng, "Emotion attribute projection for speaker recognition on emotional speech," in *Proceedings of Interspeech*, Antwerp, Belgium, Aug 2007, pp. 758 – 761.

[2] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1435–1447, 2007.

[3] L. Chen and Y. Yang, "Applying emotional factor analysis and i-vector to emotional speaker recognition," *Biometric Recognition*, pp. 174–179, 2011.

[4] I. Naseem, R. Togneri, and M. Bennamoun, "Sparse representation for speaker identification," in *Proceedings of ICPR*, Istanbul, Turkey, Aug 2010, pp. 4460–4463.

[5] M. Li, X. Zhang, Y. Yan, and S. Narayanan, "Speaker verification using sparse representations on total variability i-vectors," in *Proceedings of Interspeech*, Florence, Italy, Aug 2011, pp. 2729–2732.

[6] N. Dehak, Z.N. Karam, D.A. Reynolds, R. Dehak, W.M. Campbell, and J.R. Glass, "A channel-blind system for speaker verification," in *Proceedings of ICASSP*, Prague, Czech Republic, May 2011, pp. 4536 –4539.

[7] H. Lee, A. Battle, R. Raina, and A.Y. Ng, "Efficient sparse coding algorithms," *Advances in neural information processing systems*, vol. 19, pp. 801 – 808, 2007.

[8] I. Lopez-Moreno, C. Ortego-Resa, J. Gonzalez-Rodriguez, and D. Ramos, "Speaker dependent emotion recognition using prosodic supervectors," in *proceedings of Interspeech*, Brighton, UK, Sep 2009, pp. 1971 – 1974.

[9] Y. lin, M. Song, D. T. Quynh, Y. He, and C. Chen, "Sparse coding for flexible and robust 3d facial expression synthesis," *IEEE Computer Graphics and Applications*, vol. 32(2), pp. 76–88, 2012.

[10] T. Krech, *Die Frühsommer-meningoenzephalitis (FSME) in der Schweiz*, Ph.D. thesis, Universität Bern, 1980.

[11] T. Wu, Y. Yang, Z. Wu, and D. Li, "Masc: A speech corpus in mandarin for emotion analysis and affective speaker recognition," in *IEEE Odyssey 2006: The Speaker and Language Recognition Workshop, 2006*, Brno, Czech republic, Jun 2006, pp. 1 –5.