A RANK BASED METRIC OF ANCHOR MODELS FOR SPEAKER VERIFICATION

Yingchun YANG, Min YANG and Zhaohui WU

Zhejiang University
College of Computer Science
{yyc, rymth, wzh}@zju.edu.cn

ABSTRACT

In this paper, we present an improved method of anchor models for speaker verification. Anchor model is the method that represent a speaker by his relativity of a set of other speakers, called anchor speakers. It was firstly introduced for speaker indexing in large audio database. We suggest a rank based metric for the measurement of speaker character vectors in anchor model. Different from conventional metric methods which consider each anchor speaker equally and compare the log likelihood scores directly, in our method the relative order of anchor speakers is exploited to characterize target speaker. We have taken experiments on the YOHO database. The results show that EER of our method is 13.29% lower than that of conventional metric. Also, our method is more robust against the mismatching between test set and anchor set.

1. INTRODUCTION

Recent progresses in speaker verification have been made to model a speaker by considering its relativity to a set of reference speakers as his own character. This method was firstly introduced by by Sturim et al in 2001 to solve the problem of speaker indexing in large scale audio database [1], in which the reference speaker models are called anchor models. The set of anchor speakers forms an anchor space. Speakers are represented by Speaker Character Vectors (SCV), which indicate the speakers' projected positions in the anchor space. Then the identity of speaker can be decided based on his relative position in anchor space. Though the performance of anchor model does not reach the level of common GMM-UBM method, in situations where modelling every speaker is not feasible, like the number of speaker is too large or the data of each speaker is insufficient to build model, anchor model is proved as an effective way.

Several approaches have been done focusing on metric-based comparison of SCV. Common distance measures include the Euclidean distance [1], the vector angle [2] and the correlation [3]. There are also research works on post-processing, usually use a transformation matrix on SCV. The transformation matrix are trained from probabilistic estimation [4] or PCA/LDA orthogonalization [5].

In the anchor model method, the set of anchor speakers is a key factor. Research have been done on the influence of the number of anchor speakers on performance [2]. In this paper we focus on another issue of anchor set. In our experiment, the performance of anchor model drops significantly when the anchor set and the test speakers are mismatching. If a speaker get low scores in all anchors, which means the diversity between the speaker and all anchor speakers is great, the anchor set loses its ability as reference and can not characterize that speaker correctly. Further more, the reliability of each anchor speakers is not identical. The lower probability score a speaker gets in an anchor, the less contribution that anchor can provide in characterization of the speaker. However, in all metric-based comparisons mentioned before, this difference is not considered, and each dim in SCV is treated equally. According to that, we suggest a new distance measurement that do not directly compare the absolute scores in SCV, but compares the relative order among anchor models, the rank of each anchor.

Speaker verification experiments have been taken on the YOHO database [6]. Verification tests of a same test set is taken on two anchor sets, matching set and mismatching set. Our method is proved to be better as it get the lowest error rates in both sets. The accuracy of conventional metric decreases saliently in the mismatching set, while our method remains robust.

The next section is a brief introduction of speaker verification by anchor models. Our improved measurement is presented in section 3. Section 4 describes experiments, and section 5 analyzes the results. Section 6 gives a summary.

2. SPEAKER VERIFICATION BY ANCHOR MODELS

2.1. Anchor Model

The concept of anchor models is to represent a speaker by its relativity to a set of other speakers. It is firstly used in large audio database indexing, where the cost of building models for every speaker is unacceptable. A set of models of reference speakers, called anchor models, are trained to construct an anchor speaker space. A speaker utterance is then pro-

jected into this space by a vector constituted from scores in each anchor model. This vector characterize the speaker. It is called Speaker Character Vectors and denoted as $\hat{X}$

$$\hat{X} = \begin{pmatrix} \hat{s}(X|\bar{\lambda}_1) \\ \hat{s}(X|\bar{\lambda}_2) \\ \vdots \\ \hat{s}(X|\bar{\lambda}_n) \end{pmatrix} \tag{1}$$

in which $\hat{s}(X|\bar{\lambda}_i)$ is the average log likelihood ratio of the speaker utterance X (of N feature vectors) for the Gaussian Mixture Model of the ith anchor speaker $\bar{\lambda}_i$ relative to a Universal Background Model:

$$\hat{s}(X|\bar{\lambda}_i) = \frac{1}{N} \log \frac{p\left(X|\bar{\lambda}_i\right)}{p\left(X|\lambda_{UBM}\right)} \tag{2}$$

2.2. Verification

SCV suggests the identity of a speaker utterance. Similarity of test utterance and trained speakers is then measured by the metric of their vectors, and a threshold is used to decide whether the test utterance is spoken by the same speaker or not.

There are several metric of vectors have been studied:

- Euclidean metric [1]:

$$d(\hat{X}, \hat{Y}) = \sqrt{|\hat{X} - \hat{Y}|^2} \tag{3}$$

- Angular metric [2]:

$$\delta(\hat{X}, \hat{Y}) = \arccos\left[\frac{\hat{X}^T\hat{Y}}{\sqrt{\hat{X}^T\hat{X} \cdot \hat{Y}^T\hat{Y}}}\right] \tag{4}$$

- Correlation metric[3]:

$$\rho(\hat{X}, \hat{Y}) = 1 - \frac{C_{xy}}{\sigma_x\sigma_y} \tag{5}$$

in which $\hat{X}$ and $\hat{Y}$ are two Speaker Character Vectors, C_{xy} is the covariance between variables x, y in $\hat{X}$, $\hat{Y}$, and σ_x and σ_y are the standard deviations respectively.

In post-processing, some research works apply transformation matrixes on SCV to gain better performance. In [5], Yassine *et al* use PCA/LDA transformation to orthogonalization. Space transformation is also employed in [4] to compare the likelihood between the SCV of the test utterance and claimed speaker.

The major problem of conventional metric-based methods is the reliability of scores in SCV. The scores only suggest the degree of similarity between the speaker and the anchor speakers, but do not show how they are different. When a score gets low, which means the speaker is dissimilar to the anchor speaker, the anchor model loses its accuracy to describe that speaker. Directly comparison on scores may mistakenly contributes two different speaker to a same one, only because of they both have low scores in anchor models. The lower the probability score gets, the more possible it brings error. A uniform PCA transformation have little effort on this problem, since for different particular speakers, the principal direction or the most reliable dim of SCV is different.

3. RANK BASED METRIC

In conventional metric methods, the probability scores in SCV are compared directly and equally. Scores in SCV suggest the similarity between the speaker and anchor speakers, and are quantified as the relative log likelihood of the speaker in anchor models and the UBM. As discussed above, directly comparing on scores and considering all scores in equal fails to reflects the degree of reliability of anchor speakers.

In our opinion, though quantitive metric of scores is not proper for SCV comparison, qualitative measurement can work well. We suppose that the likelihoods of a speaker towards different anchor speakers remain in same relative levels: if a speaker gets higher score in one anchor than another, voice from the same speaker should also get higher scores in that anchor. In other words, the likelihoods of a certain speaker toward each anchor speaker is likely to keep a similar or same order.

Here we suggest a rank based comparing measure of SCV, which use the relative order of anchor models to characterize a speaker. From an SCV $\hat{X} = (s_1, s_2, \cdots, s_n)^T$, we can sort the scores s_i in a descending order:

$$s_{i_1} \geq s_{i_2} \geq \cdots \geq s_{i_n} \tag{6}$$

Then we use the rank of each anchor model to form a new vector, ranking vector $\hat{X}'$:

$$\hat{X}' = \begin{pmatrix} o_1 \\ o_2 \\ \vdots \\ o_n \end{pmatrix}, \text{ where } o_{i_j} = j, (j = 1 \ldots n) \tag{7}$$

in which i_j are the same suffixes in (6). The ranks o_i in the ranking vector are the positions in that sorted sequence. The ranking vector suggests the relative order of the similarity of the speaker to each anchor model, which is also a character of that speaker.

Instead of comparing the original SCV, we compares this ranking SCV to determine whether two speakers are identical. The Euclidean metric (3) is employed in the comparison:

$$d'(\hat{X}', \hat{Y}') = \sqrt{|\hat{X}' - \hat{Y}'|^2} = \sqrt{\sum_{i=1}^{n}(o_{xi} - o_{yi})^2} \tag{8}$$

in which o_{xi} and o_{yi} are the ranks in the ranking SCV of $\hat{X}'$ and $\hat{Y}'$ respectively.

Ranking SCV ignores details of probability scores, but focuses on the relative order of anchor speakers in anchor space, which plays an important role in characterizing the speaker's identity. Most of errors brought by low score, which are unavoidable in conventional metric methods, are eliminated in ranking SCV. Hence the performance of rank based metric should be better than conventional metric. This is proved in our experiment.

4. EXPERIMENTS

4.1. Database

Two databases are selected in evaluation of text-independent speaker verification. One is the YOHO database [6], the other is the SRMC database [7]. The YOHO database contains 108 male and 30 female speakers. It was recorded in real-world office environment, and was divided into two parts: enrollment and verification. All speakers occurs in both enrollment and verification parts. There are 4 sessions per speaker in the enrollment part, and 10 sessions per speaker in the verification. The SRMC database contains 232 male and 71 female speakers. It has 4 channels: microphone, mobile phone, PDA and telephone. Materials in each channel are further divided into several parts: personal information(PINFO), paragraph, digits, provinces and picture. In our experiment, only the microphone channel is used.

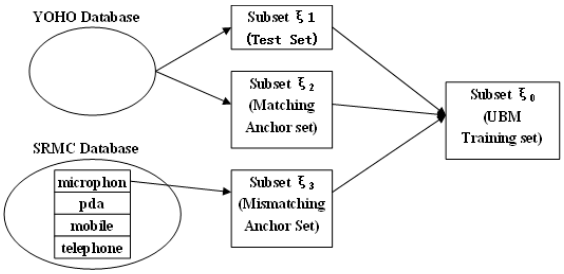

Fig. 1. Organization of experiment data sets

4 subset have been selected from utterances in these two databases, as shown in Figure 1.:

- Subset ξ_0: It includes all utterances in the enrollment part of YOHO, and all utterances in the PINFO part of microphone channel of SRMC. This subset is used to train the Universal Background Model. The total length is 21 hours approximately.
- Subset ξ_1: 50 speakers in YOHO database, 10 of which are female, are selected randomly as the evaluation subset. Both the enrollment and the verification parts of these speakers are included. Utterances in the enrollment are gathered by speaker and used as a whole to obtain the enroll SCV for each speaker. Utterances in the verify part of these speakers are used separately during the test. Each speaker has 40 utterances in the verify part. Total length in the enroll part is about 4 minutes per speaker, and utterance length in the verify part is about 5 seconds.
- Subset ξ_2: The remaining of YOHO database, 88 speakers (68 male and 20 female), are used to build the matching anchor space. One of four sessions in enrollment data (in about 1 minute) are use to train anchor models respectively.
- Subset ξ_3: 88 speakers in the SRMC database are selected randomly to build the mismatching anchor space, with the same numbers of male and female speakers as subset ξ_2. Anchor models are trained from 30 to 60 seconds long utterances in the paragraph part in this database, microphone channel.

Anchor speakers in subset ξ_2 are from the same database as the evaluation set, and there is no channel diversity. Contrastively, there is channel diversity between the mismatching subset ξ_3 and the evaluation set, which we expect to make scores in SCV generally lower than that of ξ_2. Other conditions of these two anchor sets, like the number of anchors and the length of training data, are kept to be similar. The number of speakers of each data set are relatively small and this may not reflect to merit of anchor model, but it is big enough to compare the performance of distance measurements.

4.2. Experiments Description

Feature used in all experiments are 12 dim Mel-Frequency Cepstral Coefficients (MFCC) plus energy, extracted from 32ms-long, 10ms-shifting frames. The Universal Background Model and anchor models are 64-component GMM. Anchor models are adapted from the UBM with a MAP criterion.

Six verification experiments have been taken on data set ξ_1, with the anchor model sets of ξ_2 and ξ_3, using the Euclidean metric, Angular metric and our relative distance measure respectively. No normalization of scores or distances have been applied.

5. RESULTS

Table 1. summarizes the distribution of probability scores in SCVs of both matching set and mismatching set. Scores in

Anchor Set	Mean Score	Min Score	Max Score
Matching ξ_2	-1.29	-2.43	0.78
Mismatching ξ_3	-14.69	-49.90	-5.95

Table 1. Probability score distribution in two anchor sets

the mismatching set ξ_3 is much lower than the matching set

ξ_2, that is due to channel diversity between the mismatching set ξ_3 and the test set ξ_1. There may be other reasons which also make test set and anchor set mismatching like language difference, but all reason of mismatching will reflect as low probability scores.

Figure 2. shows the performance of all three methods in both matching and mismatching set. Equal error rates (EER) are summarized in Table 2.

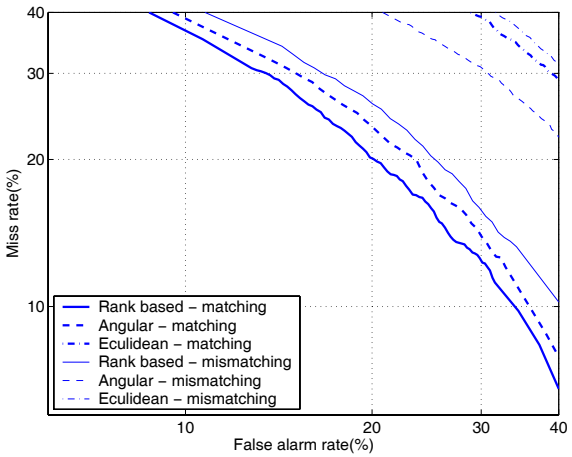

Fig. 2. DET curve for Euclidean metric, angular metric and rank based metric

Metric	Matching set	Mismatching set
Euclidean	33.25%	35.71%
Angular	21.53%	30.38%
Rank based	19.96%	22.89%

Table 2. EER for three metric in both matching and mismatching anchor sets

Generally, the EER's are higher than common GMM-UBM verification methods, that is due to the limitation of anchor model. In both matching and mismatching anchor sets, the EERs of the rank based metric are the lowest. Besides, the EERs of all methods increase in the mismatching set, while the rank based method is the most robust one.

The ranking SCV used in the rank based metric is a qualitative measurement, which only compare the relative order of anchor speakers. For different speakers, the similarity to anchor speakers are different, so the ranking SCV can characterize speaker well. Errors are brought to conventional metric which compares score equally by the different reliability level in SCV, while these error can not effect on the ranking SCV. In the mismatching case where scores are lower and the scores are less reliable, the performance of conventional metric drops more significant than rank based metric.

6. CONCLUSION

This paper presented an improved metric of anchor model. The performance of anchor model depends on many aspects. In this paper we focus on the influence of score values, and we deem the reliability of score is depend on score values. Experiments have shown that the reliability decreases when scores are low. Generally, if the speakers are more close to anchor speakers, anchor model can achieve better performance.

In respect of this fact, we proposed our rank based metric. Different from conventional metric, our method characterize the speaker not by the probability scores of anchor models, but the relative order of anchor speakers. Getting the lowest EERs in experiments with both matching and mismatching anchor sets, our method is proved better and more robust.

For future work, we will further explore the essential reason of anchor model method's limitation, and try to enhance the performance of anchor model.

7. ACKNOWLEDGEMENT

This work is supported by National Science Fund for Distinguished Young Scholars 60525202, Program for New Century Excellent Talents in University NCET-04-0545 and Key Program of Natural Science Foundation of China 60533040.

8. REFERENCES

[1] D. Sturim, D. Reynolds, E. Singer, and J. Campbell, "Speaker indexing in large audio databases using anchor models," *Proc. of ICASSP 2001*, 2001.

[2] Y. Mami and D. Charlet, "Speaker identification by location in an optimal space of anchor models," *Proc. of ICSLP 2002*, vol. 2, pp. 1333, 2002.

[3] M. Collet, D. Charlet, and F. Bimbot, "A correlation metric for speaker tracking using anchor models," *Proc. of ICASSP 2002*, 2002.

[4] M. Collet, Y. Mami, D. Charlet, and F. Bimbot, "Probabilistic anchor models approach for speaker verification," *Proc. of INTERSPEECH 2005*, pp. 2005–2008, 2005.

[5] Y. Mami and D. Charlet, "Speaker identification by anchor models with pca/lda post-processing," *Proc. of ICASSP 2003*, vol. 1, pp. 180–183, 2003.

[6] J. Campbell, "Testing with the yoho cd-rom voice verification corpus," *Proc. of ICASSP 1995*, pp. 341–344, 1995.

[7] Lifeng Sang, Zhaohui Wu, and Yingchun Yang, "Speaker recognition system in multi-channel environment," *Proc. of the IEEE International Conference on SMC 2003*, vol. 4, pp. 3116–3121, 2003.