International Conference on Communication and Signal Processing, April 3-5, 2018, India

Refining Cosine Distance Features for Robust Speaker Verification

Balasingam M D and C Santhosh Kumar

***Abstract*—Cosine distance features (CDF) have been proposed recently as a means to increase the performance of the speaker verification system. In CDF, we measure the cosine similarity of the i-vector of the input speech utterance with the i-vectors of the reference speakers. It is found that, the performance could be improved if the reference speakers are acoustically similar to the target speaker. There are different set of reference speakers for every target speakers if we select acoustically similar reference speakers, which is leading to speaker specific CDF(SSCDF).**

In this work, we explore the possibilities of further improving the performance of the SSCDF for the same number of reference speakers. We have developed two sub-systems with the reduced number of reference speakers (acoustically similar to the target speakers) and then combined/fused the decision scores of the two sub-systems. It is found that the fused SSCDF system outperformed the SSCDF consistently for reduced feature dimension and therefore using less number of reference speakers.

***Index Terms*—Speaker verification, i-vectors, Cosine Distance Features, Speaker Specific Cosine Distance Features.**

I. Introduction

Speaker verification(recognizing who is speaking) is the process of identifying the person from the charactersitics of his/her voices(voice bio-metrics). Mel-frequency cepstral coefficients (MFCC) [1] are the most popular features used in speaker verification systems. MFCC has produced considerable performance for speaker verification systems. From the recent researches, it has been noted that delta spectral cepstral coefficients(DSCC) [2], power normalized cepstral coefficients(PNCC) [3], and the mean hilbert envelope coefficients(MHEC) [4] are the popular features introduced in speaker verification systems. But, such features are not introduced as a means to increase the performance, instead for noise robustness [6].

The latest improvements in the area of speaker verification especially with cosine distance features have showed better performance for the speaker verification system. Cosine distance features (CDF) [7] [8] [9], a new feature which is extracted with the cosine similarity of the i-vector with the i-vectors of the available reference speakers. It has been proved that the CDF with back-end support vector machine classifier (CDF-SVM) performs better than i-vector with cosine distance scoring (i-CDS) [5], and an i-vector with a back-end SVM classifier (i-SVM) [10] at high SNR and noise free environments. Here we are considering all the reference speakers to develop CDF, hence this can be considered as non speaker specific CDF. Major disadvantage with the CDF is that, more the reference speakers, more the dimension. Further, CDF uses the fixed reference dictionary for all the train and test i-vectors. But the performance of the CDF system can be improved further if we select acoustically similar i-vectors for every target i-vectors while developing CDF. We obtained the speaker specific CDF(SSCDF) by considering only the acoustically similar reference speaker models for every speech utterances(training and testing). Thus every train and test i-vectors have different reference speakers i-vectors from the reference dictionary. SSCDF-SVM system has showed considerable improvement in the performance compared with normal CDF-SVM system.

In this research, We tested the performance of fusing [11] two SSCDF-SVM systems which are developed by splitting the i-vectors of the available reference speakers equally into two subset. As a part of our experiments, we have developed the i-CDS, i-SVM and CDF-SVM systems to compare the results. The experiment result shows that the performance of fusing two SSCDF-SVM systems is better than the baseline systems SSCDF-SVM, CDF-SVM, i-SVM, i-CDS.

II. System Description

To derive SSCDF for the target speaker, we consider only the acoustically similar speaker models from the reference dictionary. We have divided the available reference speakers (reference dictionary) equally into half to derive the two SSCDF sub-systems. The baseline systems CDF-SVM, i-SVM, i-CDS are also developed to compare the results.

A. i-CDS

Cosine distance scoring(CDS) is the famous technique in which the cosine distance between the test speaker i-vector and the train speaker i-vector, is considered as the decision score [5].

We can calculate the cosine distance as,

$$decision\_score(ivec_{train}, ivec_{test}) = \frac{< ivec_{train}, ivec_{test} >}{\|ivec_{train}\| \|ivec_{test}\|} \quad (1)$$

Balasingam M D and C Santhosh Kumar are with the Machine Intelligence Research Lab, Dept of Electronics and Communication Engineering, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India.
E-mail: cb.en.p2ae116006@cb.students.amirta.edu and cs_kumar@cb.amrita.edu

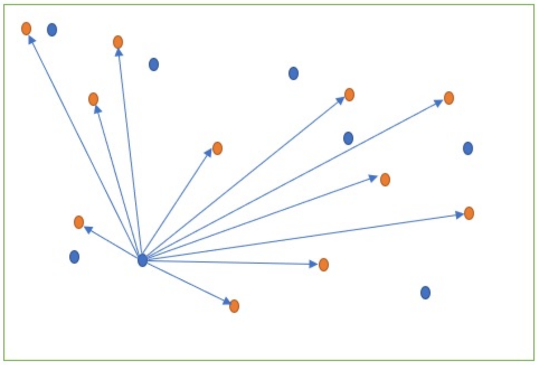

Fig. 1. Developing CDF for an i-vector where blue indicates the train and test i-vectors , red indicates reference speakers i-vector and the arrow indicates the cosine distance

where $ivec_{test}$ nd $ivec_{train}$ are the test and train i-vectors respectively.

B. i-SVM

Here, the train and test i-vectors are given to SVM(back-end) classifier to classify the positive and negative speech utterances for every target speakers. For training, usually we give background speakers speech utterances (negative samples) along with the single speech utterance of the target speaker speech utterance. Here the number of negative samples are more compared with the positive samples because of which we get considerably less performance. Utterance partitioning with acoustic vector re-sampling(UP-AVR) [12] is the technique to increase the performance of the SVM by partitioning target speaker speech utterance into many before getting i-vector. By this we can increase the number of positive samples which in turn increase the performance of SVM. In our case, we have partitioned the positive training speech utterance into 9 sub-utterances and generated the 9 i-vectors. These 9 i-vectors along with the background i-vectors are given to SVM as training data for every target.

C. CDF-SVM

Here, the cosine similarity of the i-vector with the i-vectors in the reference dictionary are considered as new features which are named as cosine distance features(CDF).

The k^{th} element of CDF for the i-vector $ivec_{input}$ can be derived as,

$$CDF(k) = \frac{< ivec_{input}, ivec_{rf}(k) >}{\|ivec_{input}\| \|ivec_{rf}(k)\|} \tag{2}$$

where $ivec_{rf}(k)$ is the k^{th} i-vector of reference speaker model. Only the speakers from the development data are considered as reference speaker models(reference dictionary).

The dimension of the CDF is determined by the number of speakers in the reference dictionary. If N reference speakers are considered, then dimension of the CDF will be N. Train and test data are given in the form of this new feature to the SVM classifier. Fig 1 shows how to derive the CDF for an i-vector.

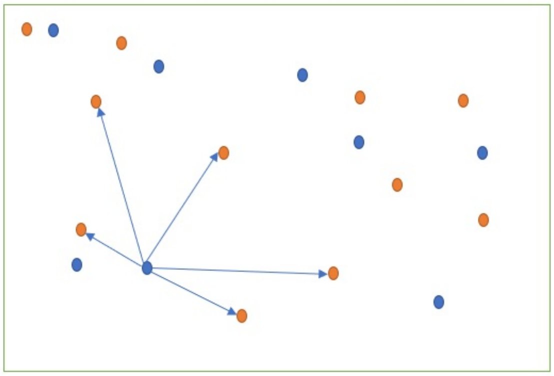

Fig. 2. Developing SSCDF for an i-vector where blue indicates the train and test i-vectors , red indicates reference speakers i-vector and the arrow indicates the cosine distance

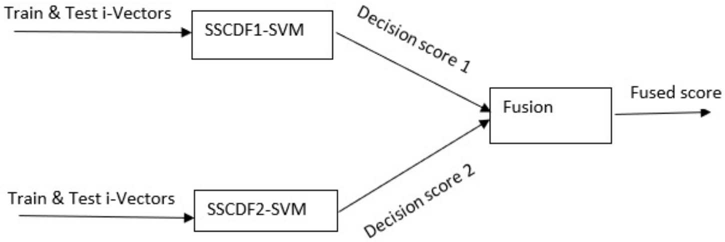

Fig. 3. Fusing two SSCDF-SVM sub-systems

D. SSCDF-SVM

In SSCDF-SVM, We consider only the acoustically similar reference speakers for every target speakers.

In Fig 2 we considered only the acoustically similar i-vectors from the reference dictionary to derive the SSCDF.

E. Fusing two SSCDF-SVM

To derive two SSCDF for an i-vector, we just split the reference dictionary into half and then develop two individual SSCDF. We have used the bosaris toolkit [10] to fuse the decision scores of two SSCDF-SVM to increase the performance. Fig 3 shows the fusion of two SSCDF.

III. Experiments and Results

For all the experiments in this research, we have used the female part of core short 2 short 3 trials of the NIST (National Institute of Standards and Technology) 2008 SRE(Speaker Recognition Evaluation) as the training and testing data. We have used NIST 2004 and Fisher part-2 as a development data. For both training and testing, we have used only the telephonic speech data. We have used spectral matching based voice activity detection [13] to remove the silence segments from the sample utterance . In this work, we have developed all the speaker verification systems with the use of short term cepstral coefficients MFCC. In MFCC extraction, 19-dimensional mel cepstral coefficients in addition with log energy, its delta and

TABLE I
PERFORMANCE COMPARISON OF I-SVM,I-CDS,CDF-SVM IN TERMS OF MINDCF

	i-SVM	i-CDS	CDF-SVM
Min-DCF	4.22	3.82	3.69

TABLE II
PERFORMANCE OF SSCDF OVER DIFFERENT DIMENSIONS

Dimensions	200	400	600	800	1000	1200	1400	1600	1800	2000
Min-DCF	3.88	3.56	3.53	3.51	3.47	3.42	3.41	**3.40**	3.42	3.48

delta delta coefficients are used to form 60 dimensional feature vector.

Universal background model (UBM) [5] is a 512-component gaussian mixture model (GMM) trained with the development data. 400 is the total variability rank, 200 is the LDA (linear discriminant analysis) rank. The total variability matrix, linear discriminant analysis and within-class co-variance normalization are trained using the development data. Then the i-vectors are derived using both the development data and evaluation data. Finally i-CDS, i-SVM, CDF-SVM, SSCDF-SVM and fusion of two SSCDF-SVM are developed.We have used Min-DCF as evaluation parameter which has been calculated from the detection error trade-off(DET) [11] curve.

Detection cost function(DCF) can be calculated as,

$$DCF(\theta) = 0.01 \times P_{miss}(\theta) + 0.99 \times P_{falsealarm}(\theta) \quad (3)$$

where $P_{miss}(\theta)$ is the miss probability and $P_{falsealarm}(\theta)$ is the false alarm probability. Min-DCF(Minimum DCF) indicates the optimum cost of the DCF value. We have used in total 3083 reference speaker models to derive the CDF. Thus the dimension of the CDF is 3083. From Table I, it is evident that performance of CDF is better than i-SVM and i-CDS.

A. SSCDF-SVM

To derive SSCDF, We have considered the acoustically similar i-vectors from the reference dictionary in numbers like 200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000 based on cosine score. For maximum dominant i-vectors from the reference dictionary we get the better performance.

From Table II, It has been shown that around 1600 number of reference speaker i-vectors we get better performance.

From Fig 4, It is clear that we have achieved better performance around 1600 reference speaker models from the given reference dictionary.

B. Fusing two SSCDF-SVM

Here, We split the reference dictionary equally into half. That is, we split the available 3083 reference speakers model into 1541 and 1542 reference speakers model. Then we derive the individual speaker specific CDF (SSCDF1, SSCDF2) for all the train and test i-vectors. Then we give the SSCDFs separately to the support vector machine(SVM) to generate the decision scores. We have used bosaris toolkit to fuse the decision scores of two sub-systems. To train the fusion system, we have used NIST 2004 SRE data.

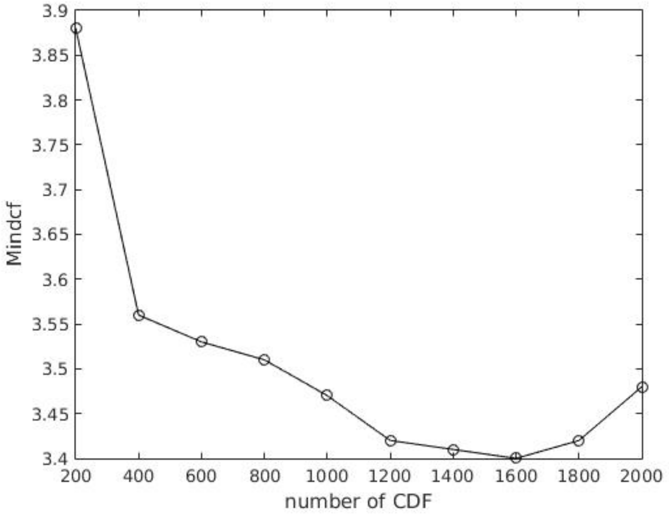

Fig. 4. Performance of SSCDF

TABLE III
PERFORMANCE OF FUSING TWO SSCDF SUB-SYSTEMS

Dimensions	200	400	600	800	1000	1200	1400	1600	1800	2000
Min-DCF	3.74	3.49	3.39	3.35	3.33	**3.32**	3.40	3.43	3.51	3.61

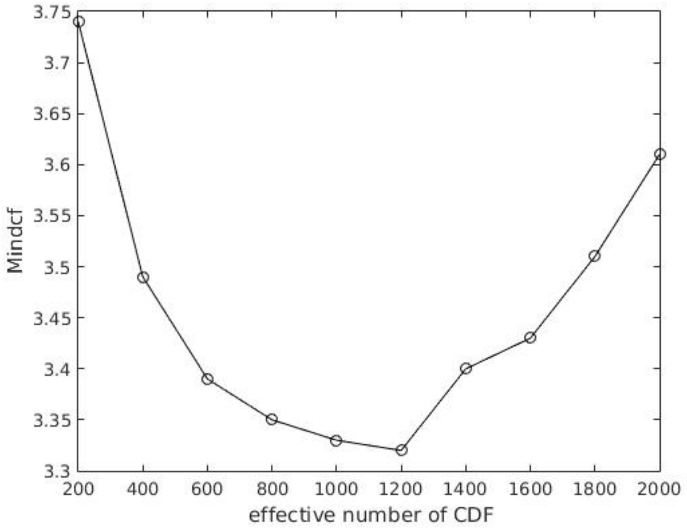

Fig. 5. Performance of SSCDF with fusion

From Table III, It has been shown that around 1200 effective reference speaker i-vectors(reference speaker i-vectors from both the system, for example if we consider 200 as the reference speaker i-vectors for a single SSCDF then the effective number of CDF is 200 + 200 ie, 400) we get better performance when fusing two SSCDF.

From Fig 5, It is clear that we have achieved improved performance with reduced effective number of reference speaker models at 1200.

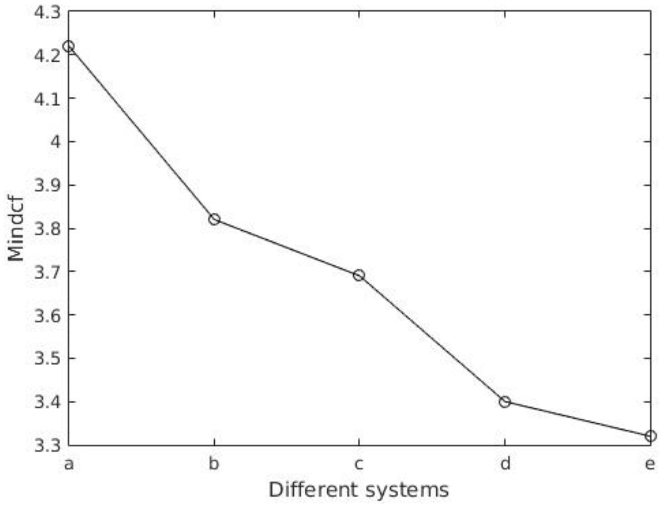

Fig. 6. Performance comparison of the proposed system(e) with the previous systems. (a) i-SVM (b) i-CDS (c) CDF-SVM (d) SSCDF-SVM (e) Fusion of two SSCDF-SVM

From Fig 6, we can see that the the proposed system (fusion of two SSCDF-SVM systems) have minimum MinDCF (3.32) which is better than all the previous systems. Thus the fused system give better performance as well as reduction in dimension than the previous systems.

IV. Conclusion

In this research, We have explored the use of fusing two speaker specific cosine distance features (SSCDF) sub-systems to increase the speaker verification system's performance in noise free environment. We have used MFCC to derive i-vectors. We have used cosine distance scoring of i-vector(i-CDS), i-vector with back-end support vector machine (i-SVM) classifier, cosine distance feature with back-end support vector machine (CDF-SVM) classifier as a baseline systems to compare the performance of SSCDF and the fused SSCDF. For all the experiments in this research, we have used the female part of core short 2 short 3 trials of the NIST 2008 SRE.

The results of the experiments show that speaker specific CDF(SSCDF) give better performance with considerable reduction in dimension, and therefore reduction in number of reference speakers. If we fuse two speaker specific CDF(SSCDF1+SSCDF2), we get further improvement in the performance with further reduction in the dimension.

Acknowledgment

Author (Balasingam M D) wishes to thank Tata Consultancy Services (TCS) for sponsoring his M.Tech. Also, author wishes to thank Kuruvachan K. George, Sreekumar K.T for their continuous support during the course of this work.

References

[1] Ezzaidi, Hassan, Jean Rouat, and Douglas O'Shaughnessy. "Combining pitch and MFCC for speaker identification systems." In 2001: A Speaker Odyssey-The Speaker Recognition Workshop. 2001.

[2] Kumar, Kshitiz, Chanwoo Kim, and Richard M. Stern. "Delta-spectral cepstral coefficients for robust speech verification." In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pp. 4784-4787. IEEE, 2011.

[3] Kim, Chanwoo, and Richard M. Stern. "Power-normalized cepstral coefficients (PNCC) for robust speech verification." In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pp. 4101-4104. IEEE, 2012.

[4] Sadjadi, Seyed Omid, Taufiq Hasan, and John HL Hansen. "Mean Hilbert envelope coefficients (MHEC) for robust speaker verification." In Thirteenth Annual Conference of the International Speech Communication Association. 2012.

[5] Dehak, Najim, Patrick J. Kenny, Rda Dehak, Pierre Dumouchel, and Pierre Ouellet. "Front-end factor analysis for speaker verification." IEEE Transactions on Audio, Speech, and Language Processing 19, no. 4 (2011): 788-798.

[6] Sarkar, S. "Robust speaker verification in noisy environments." PhD diss., Masters thesis, School of Information Technology, Indian Institute of Technology Kharagpur, 2014.

[7] George, Kuruvachan K., C. Santhosh Kumar, and Ashish Panda. "Cosine distance features for robust speaker verification." In Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[8] George, Kuruvachan K., C. Santhosh Kumar, K. I. Ramachandran, and Ashish Panda. "Cosine distance features for improved speaker verification." Electronics Letters 51, no. 12 (2015): 939-941.

[9] Kumar, C. Santhosh, Kuruvachan K. George, K. I. Ramachandran, and Ashish Panda. "Weighted cosine distance features for speaker verification." In India Conference (INDICON), 2015 Annual IEEE, pp. 1-5. IEEE, 2015.

[10] Rao, Wei, and Man-Wai Mak. "Boosting the performance of i-vector based speaker verification via utterance partitioning." IEEE Transactions on Audio, Speech, and Language Processing 21, no. 5 (2013): 1012-1022.

[11] Brmmer, Niko, and Edward De Villiers. "The bosaris toolkit: Theory, algorithms and code for surviving the new dcf." arXiv preprint arXiv:1304.2865 (2013).

[12] Mak, Man-Wai, and Wei Rao. "Utterance partitioning with acoustic vector resampling for GMMSVM speaker verification." Speech Communication 53, no. 1 (2011): 119-130.

[13] Sreekumar, K. T., Kuruvachan K. George, K. Arunraj, and C. Santhosh Kumar. "Spectral matching based voice activity detector for improved speaker verification." In Power Signals Control and Computations (EPSCICON), 2014 International Conference on, pp. 1-4. IEEE, 2014.