Forensic Applications using Cosine Distance Feature and Cepstral Coefficient for Speaker Recognition

M. Sreenivasa Reddy
Department of Mechanical Engineering, Aditya University

V. Satyanarayana
Department of ECE, Aditya University

https://doi.org/10.5109/7183442

出版情報：Evergreen. 11 (2), pp.1320-1325, 2024-06. 九州大学グリーンテクノロジー研究教育センター
バージョン：

Forensic Applications using Cosine Distance Feature and Cepstral Coefficient for Speaker Recognition

M. Sreenivasa Reddy[1], V. Satyanarayana[2*]

[1]Department of Mechanical Engineering, Aditya University, Surampalem, India

[2]Department of ECE, Aditya University, Surampalem, India

Corresponding Author E-mail: vasece_vella@adityauniversity.in

(Received September 14, 2022; Revised April 8, 2024; Accepted June 14, 2024).

Abstract: Speaker Recognition (SR) uses a person's voice to identify them. Due to their high performance and capability to recompense for session/channel inconsistencies, i-vectors have recently gained popularity as SRS input features. Additional speaker-specific perceptual cues can be derived from behaviors and learned characteristics, such as vocabulary selection, accent, intonation style, and emotional aspects. Humans also use the speaker's sound signature similarity to known speakers to improve sound recognition precision. We need a new feature vector representation that compares a mark speaker's speech to a set of reference speaker’s (codebook/dictionary). The speaker's utterance is encoded as cosine distance feature vectors (CDF). Back-end classifiers use SVMs (CDF-SVM). As a result, an SVM classifier with an intersection kernel captures the most acoustic similarities between target and reference speakers. Determining speaker discrimination is more important with reference speakers that are acoustically similar. Using CDF sparingly improves discriminative power by keeping only a few large values that correspond to the most similar reference speakers and setting all other elements to 0. On the core shorting condition of NIST's 2008 SRE databases, CDF-SVM outperforms SR systems using I-Vectors.

Keywords: Acoustic coefficients, CDF, Feature extraction, MFCC, Speaker recognition.

1. Introduction

A SR system uses unique information obtained from a person's spoken words to identify him or her automatically. Because it includes identification, authentication (verification), classification (by acoustic characteristics), segmentation, tracking, and detection, this type of speech recognition is known as voice biometrics (detection of speakers). All procedures that involve identifying someone by listening to their speech are referred to as speech recognition (SR). Facial image recognition is another important biometric identification tool, used in conjunction with fingerprints and retinal scan recognition [1, 2).

Acoustic limitations reflecting the target speaker's vocal tract characteristics are used as input features in automatic SR systems. Improved SR accuracy has also been found to be influenced by sociocultural and emotional attributes of the speaker (vocabulary selection, accent, intonation style, etc.). Besides comparing a new speaker to someone they already know; humans also look for similarities between them and the new speaker. The similarity between a target speaker and a predefined set of reference speakers (codebook/dictionary) is proposed as a feature vector for speech recognition (SR)[3)]. This chapter describes a new distance-based feature representation for SR. For each speech utterance, a feature vector is formed by computing cosine similarity between the speech utterance i-vector and the i-vectors of a set of orientation speakers (codebook/dictionary) [4)]. 'Cosine distance' is the new feature's name (CDF). Moreover, the CDF shows that reference speakers that are acoustically similar to the target speakers are further significant for speaker discernment. SR tasks are more effectively performed with a sparse representation of the CDF[5, 6)]. CDF is used as input by the SVM back-end classifier (CDF-SVM). Next, we show that the best way to capture acoustic similarity between target and reference speakers is to use an intersection kernel in the SVM classifier. In both female and male trials, SR systems based on i-vectors outperformed CDF-SVM for NIST 2008's core short2-short condition and distance measure shown in Fig. 1

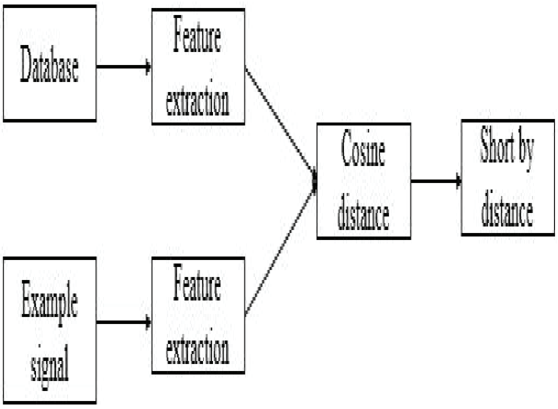

Fig. 1: Cosine distance measure for similarity measurement

These are heat maps that have been color-coded to represent the individual values within a matrix of data[7]. A variety of color schemes can be used to display heat maps. In our study, we used a red-green color map for perceptual reasons. High values are shown in red, while values in the center are shown in black. CDF is visualized using a red-green color map. Nine utterances each from five different speakers produce 61 CDF vectors. The CDF representation has five distinct vertical bands that show its speaker discrimination ability. There may be some overlap between the color maps for some reference speakers and those for target speakers at a particular CDF dimension because they are close in proximity[8, 9]. CDF vectors will be able to discriminate between target speakers because there are a large number of reference speakers in the code book/dictionary[10].

Two-dimensional (the best two feature dimensions after t-distributed stochastic neighbor implanting, or T-SNE) scatter plotted is created, and univariate histograms of the marginal distributions of the two dimensions are displayed on the horizontal and vertical axes of the scatter plot[11]. Diagrams of scatter histogram plots for i-vectors and CDF vectors are shown in Fig. 1, respectively. CDF vectors outperform i-vectors in terms of speaker discrimination. i-vectors do not reduce intra-speaker variability in any case, according to the study. In the $^{\text{fol24lowing}}$ section, we provide a more detailed analysis of the CDF representation's ability to discriminate between speakers[12].

2. Related Works

Textual communication is a natural part of the human experience. As expected, a person's voice can be used to identify him. After about 2-3 seconds of speeches, a human can recognize a speech. Studies on SR found that 97 percent accuracy was achieved when at least one sentence of speech was heard Too few or too many speakers hurt the performance. It is not uncommon for machines to outperform humans when given short test utterances and an abundance of speakers. More than 50 years have passed since the first studies of speaker identification systems were conducted [13, 14] .

The sections that follow provide a brief overview of this work. To begin speaker recognition research in 1963, Prozanski used filter banks and digital spectrograms to measure similarity. After this conversion took place, ten talkers' common phrases were converted into time, frequency, and energy patterns [15]. As part of the recognition procedure, test patterns were cross-correlated with reference patterns to determine the talker of the test utterance. Orthogonal Transforms and Vector Quantization were used to identify three-dimensional patterns and an overall recognition score of 15.89% was attained. 2008-2012 When only spectral information was retained, three-dimensional patterns were able to be recognized as exactly the same. Improving the work in [16] required a specific subset of features to be prioritized. Features were made by averaging the speech energy across rectangular spectrogram areas. To arrive at the number of features and the area used to form them, we conducted an in-depth analysis [17]. Doddington went with the formant analysis approach in the preceding two examples, instead of the filter bank approach used previously[18]. Doddington tested to see if the speaker was who they said they were by using eight speakers that he or she knew, along with 32 impersonators. A time-based method of verification was used to determine formant frequencies, voicing pitch period, and speech energy. Time normalization was found to be essential in order to improve the performance of verification errors[19].

The frequency position of formants and pitch of voiced sounds shifts lower with age over the course of 29 years. When examining the difference between the voice formant structures of the "normal" voice and the "disguised" voice, a significant difference was found[20]. A study in which the spectrograms of imitators and famous people were compared found that the spectrograms of both groups were alike features have been extracted using Mel-Frequency Cepstral Coefficients instead of LPC parameters (MFCC)[21]. Musical Frequencies-Consciousness Complexes (MFCCs) are based on the fact that human hearing's critical bandwidths change with frequency. Logarithmic and linear filters are both used in MFCC[22]. The signal is expressed on the Mel scale to capture important characteristics of speech[23]. The audio on this scale has linear frequency spacing from under 1000 Hz to under 1 KHz, and logarithmic frequency spacing from above 1 KHz to over 1 KHz. Perceptual hearing threshold at 40 decibels above the pitch of a 1 KHz tone (1000 Mel's)[24, 25] .

3. Methodology

The centroid of each speaker is calculated from multiple speech utterances made by each speaker. Using the Cosine Distance Measurement Method, additional distance measurements are calculated for a second voice sample from the same speaker The intra-speaker discrimination power of CDF representations is compared to that of i-vectors. You'll notice that in every case, the CDF representation stays closer to the centroid, which indicates that the CDF has a smaller intra-speaker variability than the i-vector. Using Euclidean distance measurements, we compare the inter-speaker discrimination capabilities of CDF representations and i-vectors. This is due to the fact that CDF centroids tend to be larger than i-vector centroids for most speakers in Fig. 2.

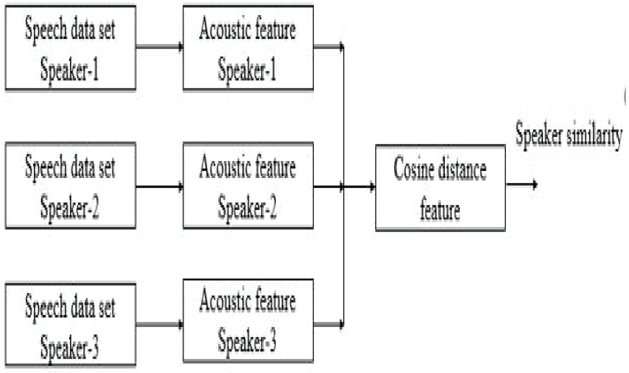
Speech data set Speaker-1
Acoustic feature Speaker-1
Speech data set Speaker-2
Acoustic feature Speaker-2
Speech data set Speaker-3
Acoustic feature Speaker-3
Cosine distance feature
Speaker similarity

Fig. 2: Speaker similarity measurement using Cosine distance feature

Each of the five languages was represented by twenty speakers. Based on the CDF vectors calculated from the 100 reference speakers, twenty reference speakers were chosen for each target utterance. Figure 2 shows a breakdown of native language speakers in the United States. 69 of the The references used were derived from the target speaker's native language. In contrast to English, languages such as Hindi, Thai, and Chinese can be examples of languages where this occurs. We had English and Vietnamese reference speakers from all over the country, but we made no special effort to gather subjects from particular locations with regard to their respective spoken languages. Indians, Chinese, Thais, and Vietnamese people migrate to the US spoke many of the English words in the database. The acoustic analysis: Mean calculated using MFCC feature extraction.

$$V_n = \{v1n,\ v2n, \dots\dots,\ vNn\};$$

$$where\ n = 1,2, \dots\dots. L \tag{1}$$

$$E_n = E(V_n)\ ;\ \ n = 1,2, \dots\dots. L \tag{2}$$

As a result, the speakers' first language had a significant impact on English, which justifies the utilization of different speakers, speaking different languages, to symbolize their primary language. Thai-speaking reference speakers were selected in greater numbers for Vietnamese speakers. This can be explained by the phonetic relationship between the two languages. Many details about the close phonetic relationship between these languages can be found in[26]. When developing SR systems, a true multilingual society requires that the first language of the target speakers be taken into account. However, the native tongue of migrants has a significant influence on the way they speak. Consider the effect of first language on enrollment utterance when selecting reference speakers for text/language independent SR based on CDF.

4. Experimental Results and Discussion

It was used as both training and test data for all experiments in the NIST SRE 2008 database. Additionally, the silence segments were removed using the VAD algorithm, which helps to make the data sets cleaner. To obtain the final feature vector, each utterance was combined with its MFCC and delta and acceleration coefficients to generate a 60-dimensional feature vector. This set of matrices was trained in the following code.

Codebooks for female and male reference speaker models were selected from the development dataset to generate CDF. It was not possible to include speakers from the training dataset due to the fact that a single target speaker had multiple speaker models associated with it The performance of i-vector and CDF systems is compared. However, CDF-SVM underperforms in male trials, while it outperforms in female trials. Fig. To test CDF-performance, SVM's we use a CDF vector with elements equal to RSC's length. In our experiments, both female and male trials have suboptimal RSC sizes, Due to the size of the codebook as a whole, the male trial codebook is significantly smaller than needed. Because of our previous tests, the number of speakers in the development databases used limited the size of the RSCs. If we could just add more RSC speakers to the female trial, we could improve the CDF-SVM system's performance a little.

By calculating the cosine distance between each speaker's centroid and the centroid of every other speaker, we are able to estimate the inter-speaker discrimination capabilities of the CDF representation and the i-vectors. The acoustic feature in terms of mean value calculated by MFCC feature[1] of 10 speakers shown in Table 1 and Fig. 3.

Table 1. Acoustic mean value of 10 speakers:

Speakers	Acoustic mean value using MFCC				
	Day 1	Day 2	Day 3	Day 4	Day 5
Speaker1	7.52	6.25	7.23	7.01	7.82
Speaker2	6.16	7.36	6.85	6.87	7.20
Speaker3	7.24	7.63	6.42	7.62	7.02

Speaker4	6.98	7.25	6.78	7.58	7.36
Speaker5	6.89	6.98	7.69	6.89	7.42
Speaker6	7.82	7.45	6.25	7.42	7.03
Speaker7	6.36	6.75	7.36	7.45	6.89
Speaker8	7.45	7.65	6.35	7.86	6.75
Speaker9	7.54	6.32	7.42	6.65	6.92
Speaker10	7.89	7.54	7.95	6.85	7.12

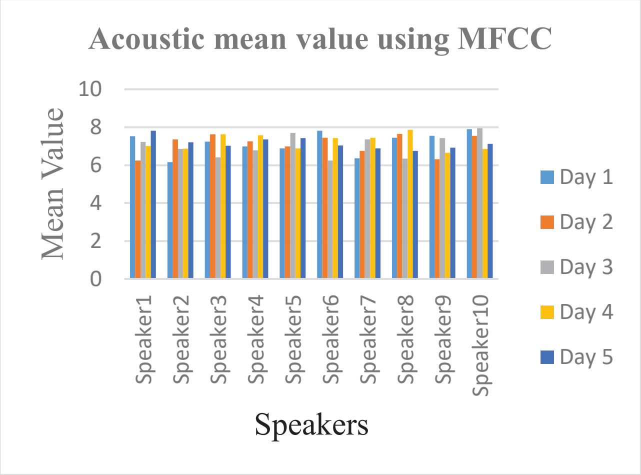

Fig. 3: Acoustic mean coefficients

Using acoustic cues from the speech increases the accuracy of speaker identification. As well as speech dynamics and tonal differences, the utterances of each speaker can be segmented using acoustic features. Speech characteristics such as pronunciation, accent, and rate of speech are influenced by the speaker's native language. People tend to use phonemes from their native language when speaking in a foreign language they have adopted. The ability to distinguish speakers based on distinctive acoustic cues in their speech will be available when working with text- and language-independent speech recognition systems. Reference speaker selection in the CDF is influenced by a speaker's native language shown in Fig. 4.

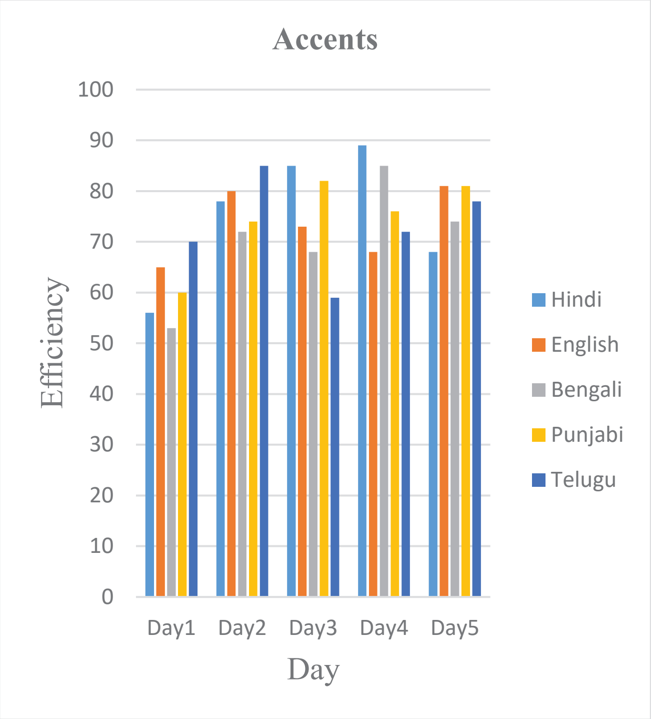

Fig. 4: Statistics on the effect of native language

For our analysis, we used twenty speakers from each of five languages: Hindi, English, Bengali, Punjabi, and Telugu. It was determined that for each target sentence the twenty most effective reference speakers would be selected based on CDF vectors generated from the 100 reference speakers.

5. Conclusion

It is proposed in this manuscript to use a distance-based SR approach. In the cosine distance feature vector, i-vectors representing reference speakers (dictionary/codebook) are added together (CDF). Ongoing research examines CDF's speaker discrimination capability and native language influence on reference speaker selection for CDF's representation. Feature-based representations performed better than SVMs with intersection kernels. Because the lower CDF elements were reduced to zeros, and because we only used reference speakers that were acoustically similar to our target speaker, we were able to achieve this result. It was found that the CDF-SVM outperformed both females and males when it came to the NIST 2008 SRE's core short2-short3 condition. EER for female speakers was increased by 0.83 percent and male speakers by 0.25 percent when compared to the best baseline system, i-PLDA. By combining CDF-SVM and i-PLDA, EER female and male trials were improved by 4.56 percent and 4.53 percent, respectively.

References

1) George, Kuruvachan K., et al. "Analysis of cosine distance features for speaker verification." *Pattern*

Recognition Letters 112 : 285-289 (2018). doi: 10.1016/ j.patrec. 2018.08.019.

2) Zou, Bei-ji, and Marie Providence Umugwaneza. "Shape-based trademark retrieval using cosine distance method." *2008 Eighth International Conference on Intelligent Systems Design and Applications*. Vol. 2. IEEE, (2008) doi: 10.1109/ISDA.2008.161.

3) Singh, Mahesh K., A. K. Singh, and Narendra Singh. "Multimedia analysis for disguised voice and classification efficiency." *Multimedia Tools and Applications* 78.(20): 29395-29411(2019). doi: 10.1007/s11042-018-6718-6.

4) Choi, Jonghyun, et al. "Toward sparse coding on cosine distance." *2014 22nd International Conference on Pattern Recognition*. IEEE, (2014). doi: 10.1109/ICPR.2014.757.

5) Singh, Mahesh K., A. K. Singh, and Narendra Singh. "Disguised voice with fast and slow speech and its acoustic analysis." *Int J Pure Appl Math* 118.(14), 241-246 (2018).doi: DOI: 10.1109/ICASSP.2018.8462169.

6) Balaji, V. Nithin, P. Bala Srinivas, and Mahesh K. Singh. "Neuromorphic advancements architecture design and its implementations technique." *Materials Today: Proceedings* (2021). doi: 10.1016/j.matpr. 2021.06.273.

7) Singh, Mahesh K., A. K. Singh, and Narendra Singh. "Multimedia utilization of non-computerized disguised voice and acoustic similarity measurement." *Multimedia Tools and Applications* 79.(47), 35537-35552 (2020). doi: 10.1007/s11042-019-08329-y.

8) Punyavathi, G., M. Neeladri, and Mahesh K. Singh. "Vehicle tracking and detection techniques using IoT." *Materials Today: Proceedings* (2021). doi: 10.1016/j.matpr. 2021.06.283.

9) Kreyssig, Florian L., and Philip C. Woodland. "Cosine-distance virtual adversarial training for semi-supervised speaker-discriminative acoustic embeddings." *arXiv preprint arXiv:2008.03756* (2020). doi: 10.1109/SLT. 2016.7846310.

10) Singh, Mahesh, Durgesh Nandan, and Sanjeev Kumar. "Statistical Analysis of Lower and Raised Pitch Voice Signal and Its Efficiency Calculation." *Traitement du Signal* 36.(5): 455-461 (2019). doi: 10.18280/ts.360511.

11) Padma, Uppalapati, Samudrala Jagadish, and Mahesh K. Singh. "Recognition of plant's leaf infection by image processing approach."*MaterialsToday: Proceedings* (2021). doi: 10.1016/j.matpr. 2021. 06.297.

12) Veerendra, G., et al. "Detecting plant Diseases, quantifying and classifying digital image processing techniques." *Materials Today: Proceedings* (2021). doi: 10.1016/j.matpr. 2021.06.271.

13) Singh, Mahesh K., Narendra Singh, and A. K. Singh. "Speaker's voice characteristics and similarity measurement using Euclidean distances." *2019 International Conference on Signal Processing and Communication (ICSC)*. IEEE, 2019.doi: 10.1109/ICSC 45622. 2019. 8938366.

14) Lei, Lei, and She Kun. "Speaker recognition using wavelet cepstral coefficient, i-vector, and cosine distance scoring and its application for forensics." *Journal of Electrical and Computer Engineering* 2016 (2016). doi: 10.1155/ 2016/ 4908412.

15) Balaji, V. Nithin, P. Bala Srinivas, and Mahesh K. Singh. "Neuromorphic advancements architecture design and its implementations technique. "*Materials Today: Proceedings* (2021). doi: 10.1016/ j.matpr. 2021.06.273.

16) Wang, Chaojun, and Fei He. "State clustering of the hot strip rolling process via kernel entropy component analysis and weighted cosine distance." *Entropy* 21.(10): 1019 (2019).doi: 10.3390/e21101019.

17) Mohd, Nik, et al. "Lattice boltzmann method for free surface impacting on vertical cylinder: A comparison with experimental data." *Evergreen: joint journal of Novel Carbon Resource Sciences & Green Asia Strategy* 4.(2): 28-37 (2017). doi: 10.5109/1929662.

18) Anushka, R. L., et al. "Lens less Cameras for Face Detection and Verification." *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)*. IEEE, 2021. doi: doi: 10.1109/ISPCC53510.2021.9609392.

19) Sharma, Manish, and Rahul Dev. "Review and Preliminary Analysis of Organic Rankine Cycle based on Turbine Inlet Temperature." *Evergreen.* 5: 22-33 (2018) doi: 10.5025/1856985.

20) Nandini, A., R. Anil Kumar, and Mahesh K. Singh. "Circuits Based on the Memristor for Fundamental Operations." *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)*. IEEE, 2021. doi: 10.1109/ISPCC53510.2021.9609439.

21) Chauhan, Shailendra Singh, and S. C. Bhaduri. "Structural analysis of a Four-bar linkage mechanism of Prosthetic knee joint using Finite Element Method." *EVERGREEN Joint Journal of Novel Carbon Resource Sciences & Green Asia Strategy* 7.02 (2020).doi: 10.5109/4055220.

22) Berawi, Mohammed Ali, et al. "Determining the Prioritized Victim of Earthquake Disaster Using Fuzzy Logic and Decision Tree Approach." *Evergreen* 7.2: 246-252 (2020).doi:10.5109/4055227

23) Akbar, Lasta Azmillah, et al. "Method development of measuring depth of burn using laser ranging in laboratory scale." *Evergreen* 7.(2) 268-274 (2020):. doi: 10.5109/4055231.

24) Yang, Haiya, and Akira Harata. "Design of a Semi-confocal Fluorescence Microscope for Observing Excitation Spectrum of Soluble Molecules Adsorbed at the Air/water Interface." *Evergreen: joint journal of Novel Carbon Resource Sciences & Green Asia Strategy* 2. (2): 1-4 (2015). doi.org/10.5109/1544074
25) Satya, P. Mohana, et al. "Stripe Noise Removal from Remote Sensing Images." *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)*. IEEE, 2021. doi: 10.1109/ISPCC53510.2021.9609457.
26) Barai, Munim K., et al. "Higher education in private universities in Bangladesh: A model for quality assurance." *Evergreen* 2.2 24-33 (2015) doi:10.5109/1544077.