A New Text Independent Speaker Recognition System with Short Utterances Using SVM

Rania Chakroun[1,3](✉) and Mondher Frikha[1,2]

[1] Advanced Technologies for Image and Signal Processing (ATISP) Research Unit, Sfax, Tunisia
chakrounrania@yahoo.fr

[2] National School of Electronics and Telecommunications of Sfax, Sakiet Ezzit, Tunisia

[3] National School of Engineering of Sfax, Sfax, Tunisia

Abstract. Recent advances in the field of speaker recognition have proved to highly outperform algorithms. However this performance degrades when limited data are presented. This paper presents examples on how Support Vector Machines (SVM) can improve speaker recognition for short utterance data duration. The main contribution in this approach is the use of new vectors when training and testing data are limited. We show how different kernels function of SVM can be used to validate the new approach with different speakers from different databases.

Keywords: Speaker recognition · Speaker verification · Speaker identification · SVM

1 Introduction

Biometric systems are essentially pattern recognition systems which operate by acquiring biometric data from an individual. Instead of the use of passwords and PIN codes which can be forgotten or stolen or using signatures which can be easily forged, body characteristics such as voice, face, fingerprints and gait have been considered as discriminative features which cannot be easily stolen or forged [1]. Human relationships are essentially based on communication between individuals. The speech in both its written and spoken form supports all aspects of human interactions. In fact, individuals can communicate with one another employing only the human vocal apparatus. Hence, the acoustic signal of human speech carries not only what is being said but also embodies individual characteristics of the speaker such as speaking styles, the speaker specific characteristics and emotions, the speaker accent, the state of health of the speaker, transmission channel properties,…etc. Every person possesses a unique voice and even when the same person says the same words, the resulting sounds can't be identical. Among the important directions in speech analysis research we find the field

R. Chakroun—No academic titles or descriptions of academic positions should be included in the addresses. The affiliations should consist of the author's institution, town, and country.

M. Themistocleous et al. (Eds.): EMCIS 2020, LNBIP 402, pp. 566–574, 2020.
https://doi.org/10.1007/978-3-030-63396-7_38

of speaker recognition. This domain has received much attention from the scientific community since many years up to the present day. Indeed, the most used in society and least importunate biometric measure is that of human speech.

In this article, we refer to speaker recognition systems which utilize human speech to recognize an individual [2]. In the past decade, numerous speaker recognition algorithms have been developed in literature [3]. However, the performances of these speaker recognition systems have usually been drastically degraded when limited data are presented.

To decrease the problem of speaker recognition based on short utterances, this article introduces a new robust speaker recognition system, which is based on new cepstral features combining between the well known state of the art Mel Frequency Cepstral Coefficients (MFCC) [3, 22] together with new robust features called Power Normalized Cepstral coefficients (PNCC) that proves to be lately efficient and successful for speech and speaker recognition applications [31–34]. We evaluate the effectiveness of these combined features on speakers taken from TIMIT [16] and VoxCeleb2 [15] databases.

The rest of this paper is organized as follows. In Sect. 2, Support Vector Machines technique is explained, Sect. 3 describe related works in speaker recognition field and explain the utility of the proposed approach, experimental protocol is presented in Sect. 4, Experimental results are demonstrated in Sect. 5 and conclusions are drawn in Sect. 6.

2 Support Vector Machines

2.1 Linear Support Vector Machines

An SVM is a classifier based on hyperplane separators. Considering the problem of separating a set of *m* training vectors S = $\{\{(x_i, y_i)\}$, where $x_i \in R^n$ is a vector of features, $y_i \in \{1, -1\}$ is a class label and i = {1,…,m}, into two different classes, with a separating hyperplane having the following equation:

$$\mathrm{wx} + \mathrm{b} = 0 \tag{1}$$

This hyperplane must maximize the margin, that's why it should satisfy the following conditions:

$$y_i(\omega.x_i) + b \geq +1 \forall i \in \{1, \ldots, m\} \tag{2}$$

The best separating hyperplane must maximize the margin M given by the equation:

$$M = \frac{2}{\|\omega\|} \tag{3}$$

In fact, the optimal hyperplane is the one that minimizes:

$$\phi(\omega) = \frac{1}{2}\omega.\omega \tag{4}$$

2.2 Non-linear Support Vector Machines

When the set of training vectors of two classes are non-linearly separable, Cortes and Vapnik [8] use new variables ξ_i to measure the miss-classification errors, with ξ_i >= 0.

For the solution of the optimisation problem, a minimization of the classification error is needed [9]. The optimal hyperplane must satisfy the following inequalities:

$$(\omega.x_i) + b \geq +1 - \xi_i, \text{ si } y_i = +1 \tag{5}$$

$$(\omega.x_i) + b \leq -1 + \xi_i, \text{ si } y_i = -1 \tag{6}$$

In this case, the optimal hyperplane is determined by the vector ω which tries to minimize the following function:

$$\phi(\omega, \xi) = \frac{1}{2}\omega.\omega + C\sum_{i=1}^{m}\xi_i \tag{7}$$

Where $\xi = (\xi_i, \ldots, \xi_m)$ and C are constants.

2.3 Kernel Support Vector Machines

When a linear boundary is inappropriate, the principle of the SVM consists in throwing the learning vectors in a high dimensional space to be able to find an optimal hyperplan.

SVM replaces the input data (x_i, x_j) with a kernel function $K(x_i, x_j)$ to constructs an optimal hyperplane in the new space. The kernel function maps the input data via an associated function Φ into a high dimensional feature space in which the mapped data can be separated linearly.

Although the existence of different kernel functions, the following functions are the most known:

- Linear: $K(x_i, x_j) = x_i^T x_j$
- Polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0.$
- Radial Basis Function (RBF): $K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2), \gamma > 0.$

Where γ, r and d are kernel parameters.

3 Related Works

For classification problems, we find that most paradigms referred to one of two families: generative models such as Gaussian Mixture Models (GMM) or discriminative classifiers like SVM. The generative models need only to train data samples from the class or target speaker and make a statistical model which describes the target speaker distribution. However, discriminative classifiers require training data for both the target and imposter speakers and generating an optimal separation between the different speakers.

Most of state-of-the-art speaker recognition systems depend on the generative training of GMM. In fact, the problem has traditionally been interpreted by directly modelling the spectral content of the speech with GMM [10]. However, the generative training of the Gaussian mixture models doesn't directly optimize the classification performance. That's why it was interesting to develop alternative discriminative approaches which address directly the classification problem [11, 12]. Some other latest works recur to the use of the neural networks technique [4]. In fact, deep neural networks (DNN) have been used for speaker verification systems [4–7].

Popular in the recent advances in speaker recognition field, the increasing adoption of SVMs, which have demonstrated to be a novel effective method for speaker recognition applications [13], [26–30]. In fact, owing to the kernel which represent the main design component in an SVM, this classifier is able to find an appropriate metric in the SVM feature space relevant to the classification problem [14]. Generally, these systems conduct to comparable or superior performances than generative methods with much less training data.

Even so, most techniques have been applied to related problems such as speaker verification, and there is a lack of effective recognition method for the short utterance text independent speaker identification task.

For speaker recognition applications, the process of feature extraction presents another fundamental phase for speaker recognitions. Indeed, this step is essential to capture the speaker specific characteristics [23]. State of the art applications use appropriate features where the most successful are the Linear Prediction Coefficients (LPCs) [17], Perceptual Linear Prediction (PLP) coefficients [20], and the latest successful and well known are spectral features which have become popular are the MFCCs Coefficients. They allow obtaining high level of performance due to the use of perceptually based Mel spaced filter bank processing of the Fourier Transform and the particular robustness to the environment and flexibility that can be achieved using cepstral analysis [3, 22].

Recently the use of the PNCC coefficients proves a great efficiency in the domain of speech recognition and also for speaker recognition applications [31–34].

In this work, we try to enhance the performance of the proposed system by using both combined MFCC and PNCC features. Thus, we profit from the robustness of both features for the task of speaker recognition. The resultant combined feature vectors are evaluated for a speaker identification system when only short utterances are available and the proposed system performance is compared against results obtained with baseline systems.

4 Experiments

4.1 Test Database

We performed our experiments using the TIMIT Dataset. The TIMIT corpus is comprised of recordings of 630 speakers (438 male, 192 female [16]) using eight major dialects of American English. Table 1 illustrates the different dialect region of TIMIT database and their respective code. For each speaker, there are ten different utterances over a clean channel. The dataset contains about 5.25 h of audio file in wav format. The sampling frequency of the utterances is 16 kHz with 16-bit resolution. The recordings are single-channel, and the mean duration of each utterance is 3.28 s.

Table 1. The different dialect regions of TIMIT database.

Dialect region	Code
New England	DR1
Northern	DR2
North Midland	DR3
South Midland	DR4
Southern	DR5
New York City	DR6
Western	DR7
Army Brat	DR8

The second set of experiments is performed using speakers from the VoxCeleb2 database [15]. This corpus contains over a million utterances from a large pool of speakers. TIMIT corpus contains clearly read speech, while VoxCeleb2 has more background noise and overlapping speech.

4.2 Acoustic Features

In our experiments, we used cepstral features extracted from the speech signal using a 25 ms Hamming window with an overlap of 10 ms. 12 MFCC Coefficients together with log energy are calculated every 10 ms. Delta and double delta coefficients were then calculated to obtain a 39-dimensional final vector. This feature vector is the most efficient in the literature [3]. We use also 39-dimensional PNCC feature vectors.

4.3 SVM Systems

The classification is realized with SVM which proved their efficiencies with regard to the other systems of classification in our domain [3, 18].

We used two SVM kernel functions in our experiments. The first one is the linear kernel. The second system uses the radial basis function kernel.

To compare our results with other approaches, we have performed two different kernel systems with low-dimensional vectors and limited training data. In fact, unlike Dehak [19] who used NIST SRE 2006 corpus where the train and test utterances contain 2.5 min of speech on average, we used utterances with a mean duration of about 3 s from TIMIT and VoxCeleb2 databases. Besides, we used MFCC features which prove their efficiency in speaker recognition [3] instead of Linear Frequency Cepstral Coefficients (LFCC) which are widely criticized because of the not linear character of the speech [21]. Moreover, as in [3], we use 39 MFCC features extracted from the speech signal instead of 60-dimensional feature vectors which are almost used in [24, 25].

Referring to the protocol suggested in [3], we use 64 speakers. For TIMIT database, we divide the utterance spoken by each speaker into 8 utterances per speaker for training and 2 utterances for testing. After that we further reduced the training duration and we use only 3 utterances per speaker for training and 2 utterances for testing. For VoxCeleb2 database, the first set of experiments is dealt with about 24 s for training and 6 s for testing. The second set of experiments is dealt with about 10 s for training and 6 s for testing.

5 Results and Discussion

We examine the performance of speaker recognition systems described previously by carrying out experimental evaluations as follows. We use two baseline systems, the first one is based on the use of MFCC features, the second baseline system is based on the use of PNCC features, and the proposed system is based on both combined MFCC and PNCC features.

The different systems for speaker recognition were implemented and evaluated with a series of experiences. For each kind of kernel, we varied its various parameters to find the values which give the optimal learning. After achieving the phase of learning, we make a set of experiences in the phase of test.

We start by presenting the first set of experiments in Table 2. For TIMIT database, we give the speaker identification rates (IR) found with linear and RBF kernels with 8 utterances per speaker for training and 2 utterances for testing. For VoxCeleb2 database, we give the results obtained with 24 s for training and 6 s for testing

From the experimental results, we notice that the use of the SVM systems with RBF kernel achieves the best identification rates.

If we compare our results to the results obtained with the baseline systems, we can remark that the proposed system outperforms the results obtained with standard MFCC coefficients and PNCC features. In fact the use of combined features allow to obtain 100% of correct identification rates against only 97.66% and 99.22% respectively with PNCC and MFCC features with TIMIT database. The results are also ameliorated for VoxCeleb2 database which attain 93.75% of correct identification rates against only 88.28% and 89.06% respectively with MFCC and PNCC features.

For further comparison, a second set of experiments was developed with shorter training duration. In fact, we use only 3 utterances for training and 2 utterances for

Table 2. Speaker identification rates with SVM-based systems using RBF and linear kernels.

Systems	TIMIT		VoxCeleb2	
	Linear kernel	RBF kernel	Linear kernel	RBF kernel
SVM baseline system with MFCC	92.18	99.22	71.88	88.28
SVM baseline system with PNCC	96.88	97.66	77.34	89.06
SVM proposed system with MFCC-PNCC	98.96	100	82.81	93.75

testing For TIMIT database and about 10 s for training and 6 s for testing with VoxCeleb2 database. The results are illustrated in Table 3.

Table 3. Speaker identification rates with SVM-based systems using RBF and Linear kernels with reduced training duration.

Systems	TIMIT		VoxCeleb2	
	Linear kernel	RBF kernel	Linear kernel	RBF kernel
SVM-baseline system with MFCC	81.25	96.09	70.31	73.44
SVM baseline system with PNCC	86.72	96.88	71.88	78.13
SVM proposed system with MFCC-PNCC	94.53	98.43	79.69	90.63

The results obtained highlight the influence of the use of short utterances in our system with limited data in the training phase. Compared to the results obtained with baseline approaches, it is clear to remark that the proposed features outperform the standard ones and allow obtaining 98.43% of correct identification rates with the RBF kernel against only 96.88% and 96.09% respectively with PNCC and MFCC coefficients. The same remark is also validated with VoxCeleb2 database which attain 90.63% of correct identification rates against only 73.44% and 78.13% respectively with MFCC and PNCC coefficients.

6 Conclusions and Perspectives

In this paper, we present a new enhanced system based on the SVM approach for speaker recognition task. This system has focused on the formulation of new features looking for recognizing speakers with much reduced information. In fact we don’t need to use additional training dataset as in traditional algorithms. Besides, we don’t require incorporating further complex algorithms. We plan the proposed features with other approaches under different conditions.

References

1. Jain, A., Ross, A., Prabhakar, S.: An introduction to biometric recognition. IEEE Trans. Circ. Syst. Video Technol. **14**(1), 4–20 (2004)
2. Reynolds, D.: An overview of automatic speaker recognition technology. In: Proceedings of IEEE International Conference Acoustics Speech Signal Processing (ICASSP), vol. 4, pp. 4072–4075 (2002)
3. Togneri, R., Pullella, D.: An overview of speaker identification: accuracy and robustness issues. In: IEEE Circuits And Systems Magazine, vol. 11, no. 2, pp. 23–61 (2011) ISSN: 1531-636X
4. Snyder, D., Ghahremani, P., Povey, D., Garcia-Romero, D., Carmiel, Y., Khudanpur, S.: Deep neural network-based speaker embeddings for end-to-end speaker verification. In: 2016 IEEE Spoken Language Technology Workshop (SLT), pp. 165–170. IEEE (December 2016)
5. Zhang, S.X, Chen, Z., Zhao, Y., Li, J., Gong, Y.: End-to-end attention based text-dependent speaker verification. arXiv preprint arXiv:1701.00562 (2017)
6. Variani, E., Lei, X., McDermott, E., Moreno, I.L., Gonzalez-Dominguez, J.: Deep neural networks for small footprint textdependent speaker verification. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 4052–4056. IEEE (2014)
7. Heigold, G., Moreno, I., Bengio, S., Shazeer, N.: End-to-endtext-dependent speaker verification. In: 2016 IEEE international conference on Acoustics, speech and signal processing (ICASSP), pp 5115–5119. IEEE (2016)
8. Cortes, C., Vapnick, V.: Support vector networks. Mach. Learn. **20**, 1–25 (1995)
9. Kamppari, S.O., Hazen, T. J.: Word and phone level acoustic confidence scoring. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (2000)
10. Reynolds, D.A., Quatieri, T.F., Dunn, R.: Speaker verification using adapted gaussian mixture models. Digital Signal Process. **10**(1–3), 19–41 (2000)
11. Keshet, J., Bengio, S.: Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods. Wiley, Hoboken (2009)
12. Louradour, J., Daoudi, K., Bach, F.: Feature space mahalanobis sequence kernels: application to svm speaker verification. IEEE Trans. Audio Speech Lang. Process. **15**(8), 2465–2475 (2007)
13. Campbell, W.M.: Generalized linear discriminant sequence kernels for speaker recognition. In: Proceedings of the International Conference on Acoustics Speech and Signal Processing. pp. 161–164 (2002)
14. Campbell, W.M., Sturim, D.E., Reynolds, D.A.: Support vector machine using GMM supervectors for speaker verification. IEEE Signal Process. Lett. **13**(5), 308–311 (2006)
15. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: deepspeaker recognition. In: Proceedings of Interspeech 2018, pp. 1086–1090 (2018)
16. Reynolds, D.A.: Automatic speaker recognition using gaussian mixture speaker models. Lincoln Lab. J. **8**(2), 173–192 (1995)
17. Atal, B.: Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J. Acoust. Soc. Am. **55**, 1304 (1974)
18. Jourani, R. Reconnaissance automatique du locuteur par des GMM à grande marge", UT3 Paul Sabatier (2012)

19. Dehak, R., Dehak, N., Kenny, P., Dumouchel, P.: Linear and non linear kernel GMM supervector machines for speaker verification. In: Proceedings of Interspeech, Antwerp, Belgium, pp. 302–305 (2007)
20. Mammone, R., Zhang, X., Ramachandran, R.: Robust speaker recognition: a feature-based approach. IEEE Signal Process. Mag. **13**(5), 58–71 (1996)
21. Pitsikalis, V., Maragos, P.: Some advances on speech analysis using generalized dimensions. In: ISCA Tutorial and Research Workshop on Non-Linear Speech Processing (NOLISP) (2003)
22. Poddar, A., Sahidullah, M., Saha, G.: Speaker verification with short utterances: a review of challenges, trends and opportunities. IET Biometrics **7**(2), 91–101 (2017)
23. Chakroun, R., Frikha, M.: Robust features for text-independent speaker recognition with short utterances. Neural Comput. Appl. **32**(17), 13863–13883 (2020). https://doi.org/10.1007/s00521-020-04793-y
24. Dehak, N., Karam, Z., Reynolds, D., Dehak, R., Campbell, W., Glass, J.: A channel-blind system for speaker verification. In: Proceedings of ICASSP, pp. 4536–4539, Prague, Czech Republic, May 2011
25. Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. **19**(4), 788–798 (2011)
26. Zhang, W.Q., Zhao, J., Zhang, W.L., et al.: Multi-scale kernels for short utterance speaker recognition. In: Proceedings of ISCSLP 2014, pp. 414–417
27. McLaren, M., Matrouf, D., Vogt, R., Bonastre, J.-F.: Applying svms and weight-based factor analysis to unsupervised adaptation for speaker verification. Comput. Speech Lang. **25**(2), 327–340 (2011)
28. Rao, W., Mak, M.W.: Construction of discriminative kernels from known and unknown non-targets for PLDA-SVM scoring. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 4012–4016. IEEE (2014 May)
29. Chakroun, R., Frikha, M.: New approach for short utterance speaker identification. IET Signal Process. **12**(7), 873–880 (2018)
30. Chakroun, R., Frikha, M.: Efficient text-independent speaker recognition with short utterances in both clean and uncontrolled environments. Multimedia Tools Appl. **79**, 21279–21298 (2020). https://doi.org/10.1007/s11042-020-08824-7
31. Kim, C., Stern, R.M.: Power-normalized cepstral coefficients (PNCC) for robust speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(7), 1315–1329 (2016)
32. Nayana, P. K., Mathew, D., Thomas, A.: Performance comparison of speaker recognition systems using GMM and i-vector methods with PNCC and RASTA PLP features. In: 2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), pp. 438–443. IEEE (2017 July)
33. Al-Kaltakchi, M.T., Woo, W.L., Dlay, S.S., Chambers, J.A.: Study of fusion strategies and exploiting the combination of MFCC and PNCC features for robust biometric speaker identification. In: 2016 4th International Conference on Biometrics and Forensics (IWBF), pp. 1–6. IEEE (March 2016)
34. Shi, X.Y., Jing, X.X., Zeng, M., Yang, H.Y.: Robust speaker recognition based on improved PNCC and i-vector. Comput. Eng. Des. **4**, 42 (2017)