International Journal of Speech Technology (2023) 26:579–587
https://doi.org/10.1007/s10772-021-09888-y

A deep learning approach for robust speaker identification using chroma energy normalized statistics and mel frequency cepstral coefficients

J. V. Thomas Abraham[1] · **A. Nayeemulla Khan**[1] · **A. Shahina**[2]

Received: 26 July 2020 / Accepted: 9 August 2021 / Published online: 30 August 2021

Abstract
Speech signals used for training and testing may vary due to mismatch in the environment or variation in the channel used or due to the physiological changes of the speaker. The performance of a speaker identification system drops significantly due to these factors. In this paper, we propose a robust speaker identification system suitable for real-world speech signal using a deep learning architecture based on convolutional neural network (CNN). Mel frequency cepstral coefficients (MFCC) features are augmented with chroma energy normalized statistics (CENS) features to train a CNN model. The VoxCeleb1 dataset is used in this study and it is found that the proposed method gives better identification accuracy than existing speaker identification methods.

Keywords Speaker identification · MFCC · Chroma features · CENS · Convolutional neural network

1 Introduction

Speech signal carries distinct information that can be used for multiple tasks such as the words spoken in a specific context (speech recognition), the person who spoke the words (speaker recognition), language used by the person to communicate the information (language identification) and who are the persons involved in the utterance (speaker diarization). Significant research progress has lead to improvements in speech, language recognition and also in producing highly robust speaker recognition systems. In speaker recognition, given an unknown speech signal, if the system determines whose speech is it, the task is known as speaker identification. Similarly the task of validating the claim of the speaker that the spoken signal belongs to himself/herself is known as speaker verification. Speaker recognition is amply used in many critical applications such as forensic speaker recognition, surveillance, authentication etc. Improving the speaker recognition system's performance is highly challenging given that the speech signals used in such applications may not be clean, and depending on the level of distortion in that signal, the system's performance may vary. The speech signal may be distorted due to the reverberation, environment changes, background noise, and babble noise, etc., Hence it is essential to build a noise-robust speaker recognition system which can perform well for speech obtained from real-world environments. Different approaches are used in building a robust speaker recognition system such as (i) enhancing the speech signal, (ii) extracting robust features from the noisy signal, or (iii) building a robust model. The methodologies evolve from simple spectrogram comparison of known and unknown signals, to model creation from extracted features and matching the extracted features of unknown signal with existing models, to state-of-the-art pattern recognition techniques.

In this paper, we propose a deep learning based approach for speaker identification. The cepstral features together with chroma features are extracted from the input signal and a convolutional neural network (CNN) model is trained with these extracted features. The CNN model is explicitly trained to discriminate between the speakers being modelled. Given an unknown speech signal, the system predicts the speaker from the set of known speaker models. We use the open source dataset, VoxCeleb for comparing our algorithm with

✉ J. V. Thomas Abraham
thomasabraham.jv@vit.ac.in

[1] VIT University Chennai Campus, Chennai, Tamil Nadu, India

[2] SSN College of Engineering, Kalavakkam, Chennai, Tamil Nadu, India

the existing ones. Our approach shows a 2% increase in the performance of the system when compared to a similar approach and other baseline systems.

This paper is organised as follows: In Sect. 2, we explain briefly a few state-of-the-art techniques related to this study; in Sect. 3 the dataset used and the experimental setup is discussed. Section 4 explains the proposed methodology and Sect. 5 discusses the results obtained.

2 Related works

The features conventionally used in speaker recognition systems can be categorized into short-term spectral features, spectral-temporal features, voice source features and high-level features. Most of the speaker recognition systems use short-term features like mel-frequency cepstral coefficients (MFCCs), linear prediction cepstral coefficients (LPCC), perceptual linear prediction (PLP) coefficients, and mel-frequency discrete wavelet coefficients (MFDWC) (Kinnunen and Li 2010). These features are more stable, need less computational power, need small amount of training data, and can be extracted easily. Early work on speaker recognition was based on using MFCC/PLP features and Gaussian mixture model (GMM) (Reynolds and Rose 1995), vector quantization (VQ) (He et al. 1997) and support vector machine (SVM) (Campbell et al. 2006) as classifiers. These approaches performed well on clean speech data and failed to perform well on noisy data. Several different approaches were used to enhance the noisy speech signal (Abraham et al. 2019; El-Fattah et al. 2014; Tavares and Coelho 2016). The combination of i-vector (Dehak et al. 2009; Kanagasundaram et al. 2011) as features and GMM-UBM as classifier, and later i-vector with PLDA (Prince and Elder 2007) had shown considerable improvement in speaker recognition. Another approach using i-vector and vector taylor series (VTS) (Lei et al. 2013) presented significant improvements in noisy conditions where noisy signals were created by artificially distorting clean speech segments with babble noises added to it.

However in this approach improvements were maintained only when the system was trained on clean data. The d-vector (Variani et al. 2014) used as feature set and DNN as a classifier showed significant improvement in recognizing the speakers. The equal error rate (EER) was found to be 14% better than the i-vector system in clean conditions and 25% better under noisy conditions. Moreover, the d-vector system was more robust to additive noise in enrollment and evaluation data.

After the rapid development of machine learning (ML) algorithms, speech / speaker recognition systems used different ML algorithms and the state-of-the-art systems use various deep learning approaches like deep neural network (DNN), CNN, recurrent neural network (RNN) (Richardson et al. 2015) etc. The features extracted from the deep learning approaches are used in GMM-UBM or i-vector based systems. All these approaches increased the computational complexity compared to traditional i-vectors due to the need for transcribed training data. In Yu et al. (2017), an adversarial network (AN) was used to generate noise robust bottleneck (BN) features and these features were used in a traditional speaker verification (SV) system. However, the number of dimensions used in this approach was 128 while in deep neural network based speech enhancement (DNN-SE), the dimension is 57. So the AN-BN front end models have a higher complexity.

The authors in Snyder et al. (2018) used an x-vector DNN embedding for robust speaker recognition by augmenting the data with reverberation and noise. This x-vector based DNN system produced a better result than the i-vector (Chang and Wang 2017) system. This approach heavily depends on large amount of data augmentation on training dataset to lower the EER and additive noises and reverberation was used for noisy signals. In Torfi et al. (2018) the authors used an adaptive feature learning technique in a 3D-CNN architecture. This model was used as a feature extractor from which speaker specific models were created. It has been observed that this system gave a better result than a d-vector method. This method extracted the spatial and temporal information jointly and captured within-speaker variations better, but it did not address about inter-speaker variability. The authors in Nagrani et al. (2017) used spectrograms and CNN and have shown that the CNN architecture performed better for both speaker verification and identification. The results were better only on pre-processing the signals using mean and variance normalization, and without any pre-processing, the results obtained were not better.

It can also be seen in literature that fusion of various types of features improves the speaker identification system performance. The authors of Lawson et al. (2011) have experimented fusion of two or three features and shown that the accuracy of speaker identification is almost always improved, with very little risk of harming accuracy. In Guo et al. (2017), cepstral coefficients like PNCC and LPCC are combined with subglottal resonances. In all noise conditions, the combined feature systems perform the best and provide relatively greater improvements for pink, factory, and low-SNR babble noise. Authors of Arsikere et al. (2014) have shown that subglottal features (SGCC) provide improved speaker recognition performance when combined with conventional MFCC features at the score level. Campbell et al. (2003) have shown that even at extremely low error rates, accuracy is improved by fusing complementary features and the information they cover is intuitively appealing. Friedland et al. (2009) have shown how the prosodic and top-ranked long-term features mainly pitch and energy dynamics can

be combined with short-term features to increase the accuracy of speaker diarization, show a consistent improvement of about 30% relative in diarization error rate compared to the best system. In a recent work, MFCC and LPC features are fused using 1D triplet CNN and it is shown that the proposed MFCC + LPC improves the SID performance on different degraded-TIMIT, Fisher and NIST SRE 2008 and 2010 datasets (Chowdhury and Ross 2020). In this paper, we have proposed fusion of MFCC with CENS features and use CNN as a classifier.

3 Datasets

The primary focus of this paper is to propose a robust speaker identification system with speech signals from uncontrolled conditions and real-world environments. The existing datasets of either clean speech signal or noisy speech signal, used in most of the speech/speaker recognition systems are created in controlled conditions. For example, speech recorded live in high quality environments such as acoustic laboratories (Millar et al. 1994; Garofolo et al. 1993), speech recorded from mobile devices or microphones (McCool et al. 2010; Woo et al. 2006) or data from telephone calls (Petrovska-Delacrétaz et al. 2000), data intercepted by police officials (Vloed et al. 2014). Some datasets though recorded in natural environment were cleaned and in Morrison and Enzinger (2016) noises and crosstalk were removed manually. For a multi-speaker recognition, datasets are taken from recorded meeting data (Janin et al. 2003; Mccowan et al. 2005) or from audio broadcasts (Bell et al. 2015) that contains audio samples under less controlled conditions. Some datasets like NTIMIT, CTIMIT and NIST SRE (Garofolo et al. 1993) are artificially degraded to represent a real world noise and are used in several speaker recognition systems (SRS). These datasets lack real world conditions.

In order to develop a speaker identification (SI) system for real world environments, we have used the VoxCeleb1 (Nagrani et al. 2017) dataset containing more than 146k utterances of 1251 celebrities, extracted from YouTube videos, shot in a large number of challenging multi-speaker acoustic environments. These include outdoor stadium, excerpts from multimedia that were professionally shot, red carpet interviews, speeches addressed to large audiences, videos shot on hand-held devices, and interviews taken in studios. As the speech is taken from day-to-day lively activities, the speech signals are degraded with real world noise, consisting of overlapping speech, background laughter, chatter, room acoustics, and channel noise. The dataset consists of approximately 55% male speakers and 45% female speakers. Speakers are taken from diverse ethnicities, accents, ages and professions. For our speaker identification task, we used 1251 speakers from the training and test set combined and we split the dataset into training, development and test sets in the ratio of 80:10:10 respectively (Table 1).

4 Proposed methodology

As described in Sect. 2, most of the CNN based systems use raw speech signal to learn features, model the speakers and then to recognize the test speech signal. In this study, we propose the use of CNN architecture to operate on combined speech features comprising of mel frequency cepstral coefficients and chroma energy normalized statistics, instead of the raw signal, for modeling higher level representation of speaker-specific characteristics.

4.1 Feature extraction

The mel-frequency coefficients derived from the speech signal model the human perception of the frequency content in the signal on the mel-scale. The non-linear mel-scale of mel-frequency is used to compute a log power spectrum of the short term power spectrum of the speech, which is then converted to a mel frequency cepstrum (MFC) by applying a linear cosine transformation. The coefficients collectively making up the MFC are known as the MFCCs. Since the mel-scale closely relates to the human auditory system better than linearly spaced frequency bands, MFCC features are used in most speaker recognition systems.

Different types of acoustic signals like environmental sounds, speech or music, each possess specific sound characteristics, and the differences in their characteristics can be detected both in the time and frequency domains. The differences can also be detected in the structure and the semantics of the signal. It is known that speech and music signals have a harmonic structure in their spectra (Alias et al. 2016).

The spectral energy distribution abridged into 12 semitones across octaves on an equal-tempered scale within a short frame is called the Chroma feature. Chroma captures the pitch class distribution present in the input signal. The Chroma feature can be extracted using different approaches. In the approach followed, the input signal is first represented in its frequency domain and the main frequency components are identified. These frequency components are mapped onto

Table 1 VoxCeleb1 dataset distribution

Number of speakers	1251
Number of male speakers	690
Number of videos/speaker	36/18/8
Number of utterances/speaker	250/123/45

Numbers in the three entries represent maximum/average/minimum

pitch classes based on their estimated tuning frequency. The vector of Chroma features is obtained from the given interval resolution and then post-processed to get the final Chroma representation. A chromogram is a spectrum based energy representation of 12 pitch classes within an octave and it is computed from a logarithmic STFT (Bartsch and Wakefield 2005). A variation of chroma known as chroma energy normalized statistics (CENS) is more robust to temporal and timbre variations (Müller et al. 2005). CENS features are computed by applying a quantization and smoothing over Chroma vectors and by an optional downsampling. Here we do not perform any downsampling and keep the original sampling rate of the input speech signals.

In the literature, Chroma features are mainly used in music analysis. In Sell and Clark (2014) authors have used chroma features to discriminate speech and obtained 97.1% precision. Since pitch is one of the features used to discriminate speakers, and Chroma features capture the pitch class, we propose the fusion of CENS features with MFCC features serving as input to our CNN model for speaker identification.

The wave files are pre-emphasized first and segmented into frames of size 50 ms with a shift of 25 ms, such that the frame length overlaps by 50%. For each of these short-term frames, we compute 40-dimension mel frequency cepstral coefficients, their delta coefficients and 12 chroma energy normalized statistics features. These features are concatenated at frame level that results in a sequence of n short-term feature vectors each of size 92 coefficients for the whole signal, where n is the number of frames in the signal. Let $c_t = [c_0, c_1, c_2, \ldots, c_{39}], t = 1 \ldots, n$ represents the MFCC features of the speech signal at time frame t, $d_t = [d_0, d_1, d_2, \ldots, d_{39}], t = 1 \ldots, n$ represent the first order derivative of c_t and $p_t = [p_1, p_2, \ldots, p_{12}], t = 1 \ldots, n$ represent the CENS features at frame index t. These first order derivatives and CENS features are concatenated with the original MFCC coefficients to yield an augmented feature vector (F_t), $F_t = [c_0, c_1, c_2, \ldots, c_{39}, d_0, d_1, d_2, \ldots, d_{39}, p_1, p_2, \ldots, p_{12}]$, $t = 1 \ldots, n$. A visual representation of the 40-dimension MFCC features, 12 dimension CENS features and the fusion of MFCC, its first-order derivative and CENS is shown in Fig. 1.

4.2 CNN model and architecture

Convolutional neural network is a type of deep learning algorithm that performs well for images. A CNN learns feature representations from a set of training images assigning weights based on their importance. With the help of relevant

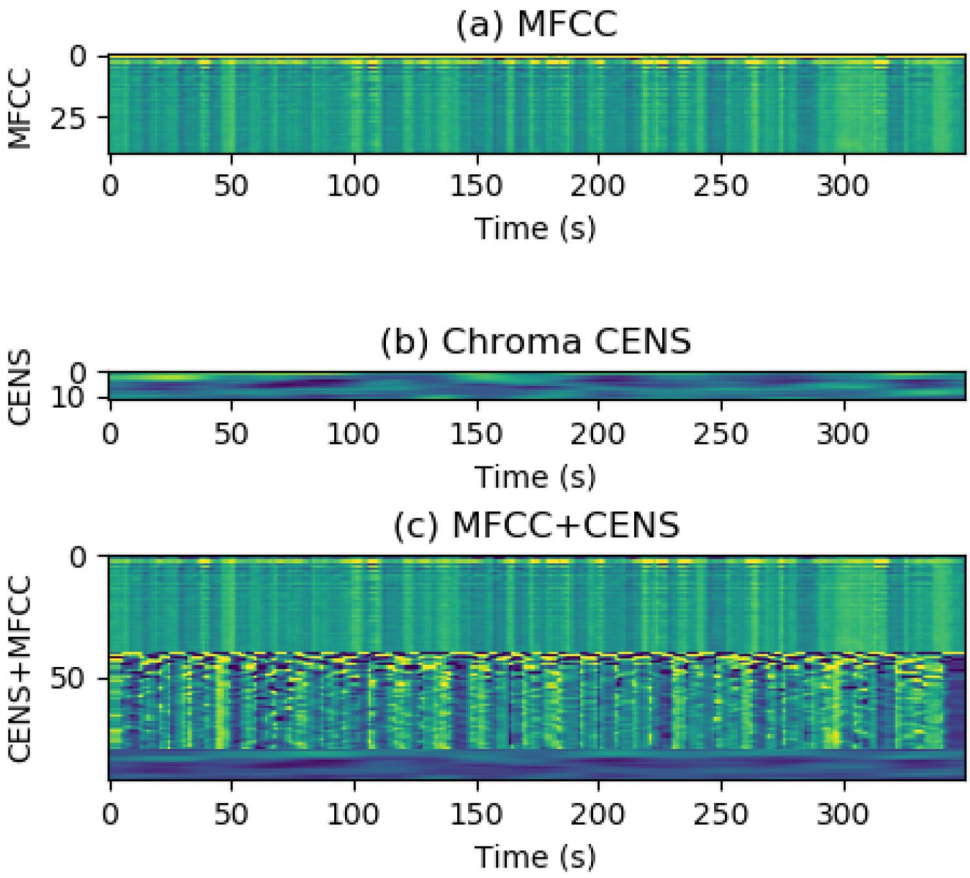

Fig. 1 Visual representation of the features **a** MFCC features **b** CENS features **c** Fusion of MFCC+delta-MFCC+CENS features

filters, a CNN captures the temporal and spatial dependencies in an image enabling better classification.

A typical CNN architecture consists of one or more convolutional and pooling layers followed by fully connected layers as shown in Fig. 2. A convolution layer with kernels extracts features from the input image. While the initial layer extracts low-level features, the layers at deeper levels extract high-level features.

The size of the convolved feature is reduced using pooling layer. Moreover, the pooling layer is used to extract the governing features, to make the model training more effective. The combination of convolutional layer and the pooling layer, together form the ith layer of a CNN. The combination of these layers may be increased depending on the complexities in the images. A fully connected layer acts like a neural network which gets a flattened input (ie. the feature map matrix will be converted as a vector). Lastly, an activation function such as softmax or sigmoid is used in the final layer to classify the outputs. We have considered VGG16, ResNet50 and Inception-ResNet v2 as three base CNN models that are explained briefly below. We have also used a custom architecture for the task.

4.3 VGG16

VGG16 is the convolution neural network (CNN or ConvNet) proposed for large-scale identification of images. It was trained on ImageNet dataset, which is a compilation of over 14 million photographs from over 22,000 classes. There are 5 Convolution blocks and 1 fully connected block in VGG16. There are several convolution layers and 1 pooling layer in each convolution block. Convolution layers use filters of size 3×3 and 2×2 Max Pooling is used for pooling layers. There are 3 FC layers in the fully connected block, the first and second have 4096 units while the third has 1000 units based on the number of classes in the ImageNet (Simonyan and Zisserman 2014).

4.4 ResNet50

Experimental SetupResNet50 architecture implemented the idea of Residual Network to address the problem of the vanishing/exploding gradient, which uses a technique called skip connections and it is the core of the residual block. ResNet network uses a VGG-19 inspired 34-layer plain network architecture in which the shortcut connection is then inserted. The architecture then transforms these shortcut connections into a residual network. ResNet50 model has 48 Convolution layers along with 1 MaxPool and 1 Average (Pool layer He et al. 2016).

4.5 InceptionResNetV2

Inception-ResNet v2 is a convolutional neural architecture based on the Inception family of architectures but includes residual connections (replacing the filter concatenation stage of the Inception architecture). Inception-ResNet-v2 is a modification of Inception v3 model, which is much deeper than the previous Inception V3. The network is 164 layers deep and can categorize pictures into 1000 types of objects and it is found more accurate than previous state-of-the-art models for image classification (Szegedy et al. 2016).

4.6 Experimental setup

As explained in Sect. 4.1 we extract MFCC and CENS features. We have not used any pre-processing techniques like voice activity detection, silence removal or unvoiced speech removal on the speech signals. The MFCC and CENS vectors extracted differs in size for different speech input,

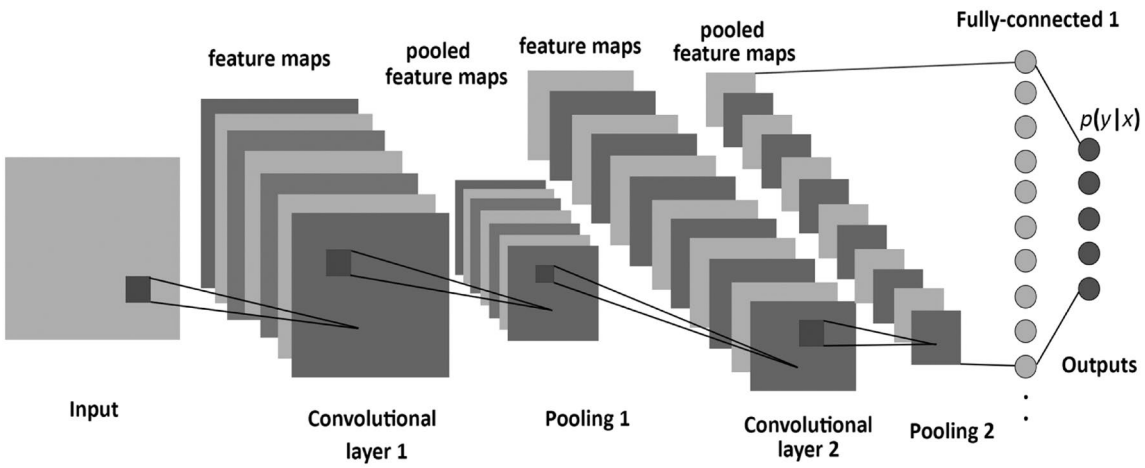

Fig. 2 Typical CNN architecture medium (Convolutional Neural Networks 2018)

and since the ConvNets can't deal sequence data of varying size, we have fixed the vector size for all speech files. We consider the first 196 frames of each input file. The 40 dimension MFCC features with its first order derivatives and 12 dimension CENS features together results in 92×196 matrix where 92 is the number of feature dimension and 196 is the number of frames. For certain speech files with less than 196 frames, we pad with a constant value 0.

Let $X_i = \{X_1, X_2, \ldots, X_n\}$ be the training set of n speakers where each speaker X_i has m speech utterances. Let $Y = \{y_1, y_2, \ldots, y_n\}$, be the label set, where $y_i \in \{0, 1, \ldots, 1251\}, 1 \leq i \leq n$ indicates the speaker identifier. We first train the CNN as a classifier such that $f(X) \approx Y$ and the model was trained with 100 epochs. Later, given an unknown speaker signal, the combined MFCC and CENS features are extracted and the previously trained classifier is used to identify the speaker of the unknown signal. The overall block diagram of the speaker identification system is given in Fig. 3.

The proposed CNN model consists of five convolution layers and two fully connected layers. The block diagram of the proposed five-layer CNN is shown in Fig. 4.

Each convolution layer has a 3×3 kernel and the number of filters used in the five convolution layers were 16, 32, 64, 128 and 256 respectively. Rectified linear unit (ReLU) is used as the activation function in all the five layers. Each convolution layer is followed by a maxpool layer with a pool size 2 and stride value 2 to reduce dimensionality of the feature maps and then batch-normalization was performed. We then flatten to transform a two-dimensional matrix of features into a vector which is then fed into a dense layer. The first dense layer, comprises of number-of-classes $\times$ 2 nodes with ReLU as an activation function. We used L2 regularization at this hidden fully connected layer to deal with the variance observed during training sessions and we chose to add a dropout layer which is useful to prevent from overfitting with a rate as 0.25. Finally, we added the last dense layer, composed of nodes equaling the number-of-classes with softmax as an activation function, which will be the output layer of the CNN.

5 Results and discussion

In this section, the performance accuracy obtained by our proposed feature set and CNN model is compared to a number of state-of-the-art methods on VoxCeleb1. The experiment was conducted with VoxCeleb1 dataset consisting of 1251 speakers. Initially a CNN model was constructed from scratch with 3 convolutional layers followed by a fully

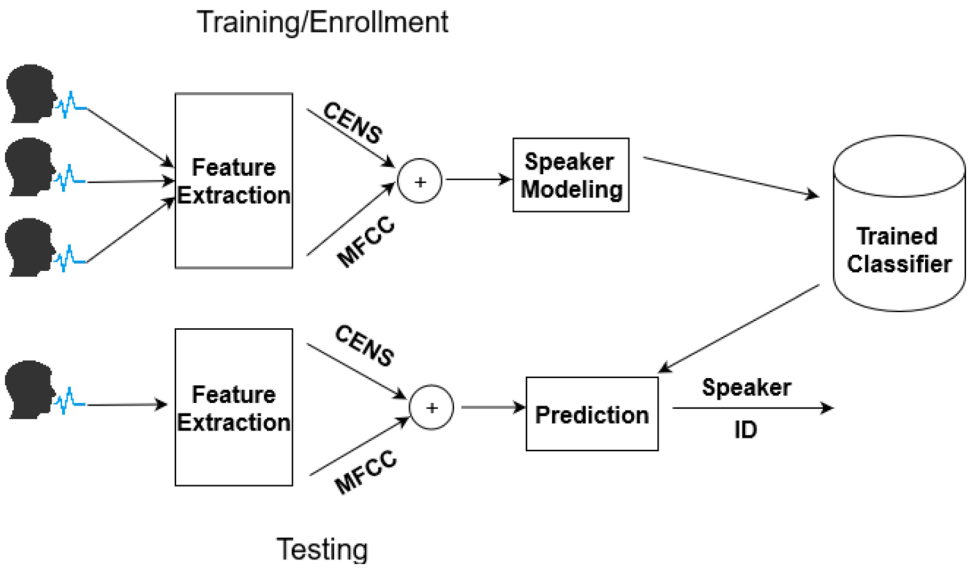

Fig. 3 Block diagram of the proposed speaker identification system

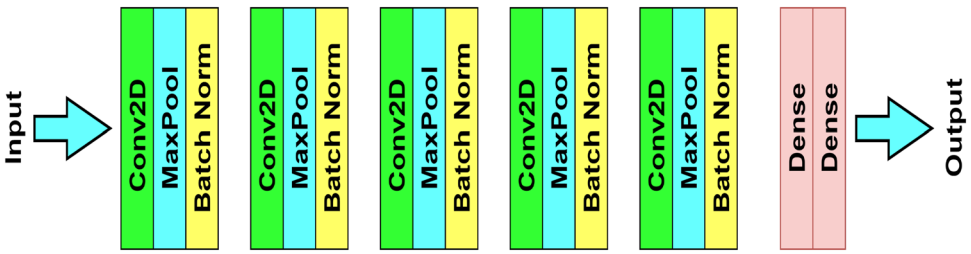

Fig. 4 Proposed CNN model

Table 2 Model hyperparameters

CNN Layers	FC Layers	Batch size	Optimizers	Activations	Dropout
5	2	128	Adam	ReLu + Softmax	0.25

connected layer. The model had high training loss and error rate. The experiments were repeated with more layers and depth to reduce the training and validation loss. In order to optimize the parameters we performed a grid search with different parameters like: tanh and relu for activation functions; adadelta, adam and rmsprop for optimizer; 0.1, 0.2 and 0.25 for dropouts; The optimal hyperparameter chosen is shown in Table 2.

Once the model parameters are optimized, the model was trained and evaluated for 1251 speakers. The experiments are conducted in an approach similar to that reported by the developer of VoxCeleb dataset. The training and the testing are performed on the same speaker. The output of the last fully connected layer is given to the softmax layer whose cardinality is N = 1251, where N is the number of speakers. The model was trained with approximately 122,000 samples for 100 epochs. We repeated the experiments several time and shuffled the data each time, ensuring that the training and test set are different at each experiment. Initially, the CNN model was trained using only MFCC features and an identification rate of 80.27% was obtained. The experiment was repeated with CENS features augmented with MFCC features and it resulted in 82.62% identification accuracy. The prosodic feature, pitch, is one of the features used to discriminate speakers, and since Chroma features capture the pitch class, augmentation of CENS features with MFCC results in the improvement of identification accuracy. For identification, the margin over the baselines is narrower, but still a significant improvement is obtained, with concatenation of CENS being the crucial step.

Results obtained for speaker identification with optimal parameter values is given in Table 3. The traditional methods like i-Vectors and SVM or i-Vector/PLDA and SVM failed to perform well on this real-time speech signal dataset. The authors in Nagrani et al. (2017) proposed CNN architecture with three variants: (i) CNN with variable length test data (ii) CNN with 3 s test data variance normalized and (iii) with no normalization. In all three experiments, a significant improvement in identification accuracy was obtained. Our proposed method using MFCC features augmented with CENS features gives Top-1 and Top-5 accuracy of 82.62% and 93.22% which is almost 23% higher than other traditional methods and 2% higher than other CNN based approaches. Figure 5 shows the Top-1 and Top-5 accuracy on the training and validation dataset of the proposed methodology.

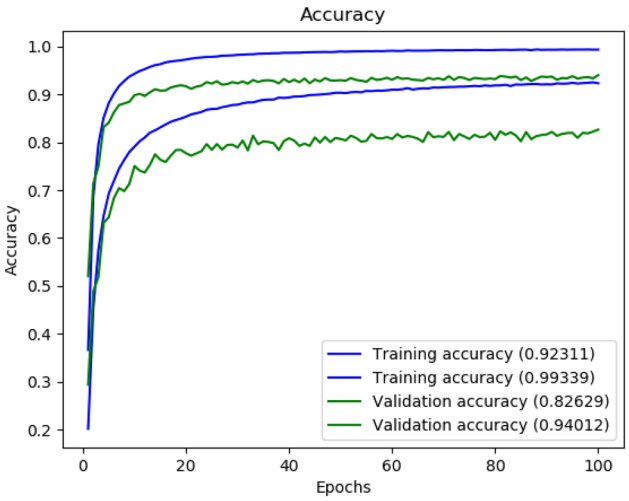

Fig. 5 Top-1 and Top-5 accuracy on training and validation dataset

Table 3 Speaker identification accuracy on test dataset

	Top-1 accuracy (%)	Top-5 accuracy (%)
i-vector + SVM (Nagrani et al. 2017)	49.0	56.6
i-vector + PLDA + SVM (Nagrani et al. 2017)	60.8	75.6
CNN-fc-3s no var. norm. (Nagrani et al. 2017)	63.5	80.3
CNN-fc-3s (Nagrani et al. 2017)	72.4	87.4
CNN (Nagrani et al. 2017)	80.5	92.1
MFCC embedding + proposed CNN	80.27	90.52
MFCC + CENS embedding + proposed CNN	82.62	93.22

Table 4 Comparison of our model with other state of the art models

CNN model	Accuracy (%)
VGG16	56.6
ResNet50	74.21
InceptionResNetV2	74.57
Proposed CNN	82.62

We also compare performance of our custom CNN architecture to a number of other state-of-the-art deep learning methods. We use the same MFCC + CENS features and only vary the network architecture, i.e., we compare against three models (VGG16, ResNet-50, InceptionResNetV2). In Table 4,

the identification accuracy of our model is compared with the recent models and our proposed CNN model achieves comparable performance on the VoxCeleb1 dataset.

6 Conclusion

Most of the existing speaker recognition systems use simulated noisy signal to build a robust system. In this study, we have considered a large scale real-world noisy speech signals recorded in different environments, and we have proposed the fusion of the mel-frequecy cepstral coefficient (MFCC) features with chroma energy normalized statistics (CENS) features, a variant of chroma features. The CENS features are mostly used in music synthesis. As CENS features capture the pitch classes of the audio signal, we extract the CENS features from speech signal and MFCC features are augmented with CENS features. It is observed that our proposed CNN model with optimized hyperparameters trained with MFCC-CENS gives better identification accuracy over other state-of-the-art methods on the real world dataset.

References

Abraham, J. V. T., Shahina, A., & Khan, A. N. (2019). Enhancing noisy speech using WEMD. *International Journal of Recent Technology and Engineering, 7,* 705–708.

Alias, F., Carrié, J. C., & Sevillano, X. (2016). A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. *Applied Sciences, 6,* 143.

Arsikere, H., An, H., & Alwan, A. (2014). Speaker recognition via fusion of subglottal features and MFCCs. In INTERSPEECH 2014.

Bartsch, M., & Wakefield, G. (2005). Audio thumbnailing of popular music using chroma-based representations. *IEEE Transactions on Multimedia, 7,* 96–104.

Bell, P., Gales, M. J. F., Hain, T., Kilgour, J., Lanchantin, P., Liu, X., McParland, A., Renals, S., Saz, O., Wester, M., & Woodland, P. C. (2015). The MGB challenge: Evaluating multi-genre broadcast media recognition. In IEEE workshop on automatic speech recognition and understanding (ASRU) (pp. 687–693).

Campbell, J., Reynolds, D., & Dunn, R. (2003). Fusing high- and low-level features for speaker recognition. In In INTERSPEECH (pp. 2665–2668).

Campbell, W., Campbell, J., Reynolds, D., Singer, E., & Torres-Carrasquillo, P. (2006). Support vector machines for speaker and language recognition. *Computer Speech & Language, 20,* 210–229.

Chang, J., & Wang, D. (2017). Robust speaker recognition based on DNN/i-vectors and speech separation. In IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 5415–5419).

Chowdhury, A., & Ross, A. (2020). Fusing mfcc and lpc features using 1d triplet cnn for speaker recognition in severely degraded audio signals. *IEEE Transactions on Information Forensics and Security, 15,* 1616–1629.

Convolutional Neural Networks. (2018). https://www.datasciencecentral.com/profiles/blogs/understanding-neural-networks-from-neuron-to-rnn-cnn-and-deep.

Dehak, N., Dehak, R., Kenny, P., Brummer, N., Ouellet, P., & Dumouchel, P. (2009). Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. Proceedings of the annual conference of the international speech communication association, INTERSPEECH (vol. 1, pp. 1559–1562).

El-Fattah, M. A. A., Dessouky, M. I., Abbas, A. M., Diab, S. M., El-Rabaie, E.-S.M., Al-Nuaimy, W., et al. (2014). Speech enhancement with an adaptive wiener filter. *International Journal of Speech Technology, 17*(1), 53–64.

Friedland, G., Vinyals, O., Huang, C., & Müller, C. (2009). Fusing short term and long term features for improved speaker diarization. In IEEE international conference on acoustics, speech and signal processing (pp. 4077–4080).

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., & Pallett, D. (1993). DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1–1.1. *NASA STI/Recon Technical Report, 93,* 27403.

Guo, J., Yang, R., Arsikere, H., & Alwan, A. (2017). Robust speaker identification via fusion of subglottal resonances and cepstral features. *The Journal of the Acoustical Society of America, 141*(4), EL420–EL426.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In IEEE conference on computer vision and pattern recognition (CVPR) (pp. 770–778).

He, J., Liu, L., & Palm, G. (1997). A new codebook training algorithm for VQ-based speaker recognition. *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2,* 1091–1094.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., & Wooters, C. (2003). The ICSI meeting corpus. In IEEE international conference on acoustics, speech, and signal processing (vol. 1).

Kanagasundaram, A., Vogt, R., Dean, D., Sridharan, S., & Mason, M. (2011). i-vector based speaker recognition on short utterances. In Proceedings of the annual conference of the international speech communication association, INTERSPEECH.

Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication, 52,* 12–40.

Lawson, A., Vabishchevich, P., Huggins, M., Ardis, P., Battles, B., & Stauffer, A. (2011) Survey and evaluation of acoustic features for speaker recognition. In IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 5444–5447).

Lei, Y., Burget, L., & Scheffer, N. (2013). A noise robust i-vector extractor using vector taylor series for speaker recognition. In IEEE international conference on acoustics, speech and signal processing (pp. 6788–6791).

McCool, C., Marcel, S., & "MOBIO Database for the ICPR . (2010). Face and Speech Competition. Idiap-Com Idiap-Com-02-2009. Idiap, 11, 2009.

Mccowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kronenthal, M., Lathoud, G., Lincoln, M., Masson, A. Lisowska., Post, W., Reidsma, D., & Wellner, P. (2005). The AMI meeting corpus. In International conference on methods and techniques in behavioral research.

Millar, J. B., Vonwiller, J. P., Harrington, J. M., & Dermody, P. J. (1994). "The Australian National Database of Spoken Language. In Proceedings of IEEE international conference on acoustics, speech and signal processing (vol. i, pp. I/97–I100).

Morrison, G. S., & Enzinger, E. (2016). Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting

those of a real forensic case (forensic_eval_01) introduction. *Speech Communication, 85,* 119–126.

Müller, M., Kurth, F., & Clausen, M. (2005). Audio matching via chroma-based statistical features. In 6th International conference on music information retrieval, ISMIR (pp. 288–295).

Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb: a large-scale speaker identification dataset. In INTERSPEECH.

Petrovska-Delacrétaz, D., Hennebert, J., Melin, H., & Genoud, D. (June 2000). POLYCOST: A telephone-speech database for speaker recognition. *Speech Communication, 31,* 265–270.

Prince, S. J. D., & Elder, J. H. (2007). Probabilistic linear discriminant analysis for inferences about identity. In IEEE 11th international conference on computer vision (pp. 1–8).

Reynolds, D., & Rose, R. (1995). Robust text-independent speaker identification using Gaussian Mixture speaker models. *IEEE Transactions on Speech and Audio Processing, 3,* 72–83.

Richardson, F., Reynolds, D., & Dehak, N. (2015). Deep neural network approaches to speaker and language recognition. *IEEE Signal Processing Letters, 22,* 1–1.

Sell, G., & Clark, P. (2014). Music tonality features for speech/music discrimination. In IEEE international conference on acoustics (pp. 2489–2493). ICASSP: Speech and Signal Processing—Proceedings.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust DNN embeddings for speaker recognition. In IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 5329–5333).

Szegedy, C., Ioffe, S., & Vanhoucke, V. (2016). Inception-v4, inception-resnet and the impact of residual connections on learning. CoRR, vol. abs/1602.07261.

Tavares, R., & Coelho, R. (2016). Speech enhancement with nonstationary acoustic noise detection in time domain. *IEEE Signal Processing Letters, 23*(1), 6–10.

Torfi, A., Dawson, J., & Nasrabadi, N. M.(2018). Text-independent speaker verification using 3D Convolutional Neural Networks. In IEEE international conference on multimedia and expo (ICME) (pp. 1–6).

Variani, E., Lei, X., McDermott, E., Moreno, I. L., Gonzalez-Dominguez, J. (2014). Deep neural networks for small footprint text-dependent speaker verification. In IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 4052–4056).

Vloed, D. van der., Bouten, J., & Leeuwen, D. Van. (2014). NFI-FRITS: A forensic speaker recognition database and some first experiments. In Proceedings of Odyssey speaker and language recognition workshop (pp. 6–13).

Woo, R. H., Park, A., & Hazen, T. J. (2006). The MIT mobile device speaker verification corpus: Data collection and preliminary experiments. In IEEE Odyssey—the speaker and language recognition workshop (pp. 1–6).

Yu, H., Tan, Z.-H., Ma, Z., & Guo, J. (2017). Adversarial network bottleneck features for noise robust speaker verification. In INTERSPEECH (pp. 1492–1496).