Hybrid Network For End-To-End Text-Independent Speaker Identification

Wajdi Ghezaiel*, Luc Brun† and Olivier LÉZORAY†
*Normandie Université, UNICAEN, ENSICAEN, CNRS, NormaSTIC, Caen France
† Normandie Université, UNICAEN, ENSICAEN CNRS, GREYC Caen, France
Email: wajdi.ghezaiel@ensicaen.fr, luc.brun@ensicaen.fr, olivier.lezoray@unicaen.fr

***Abstract*—Deep learning has recently improved the performance of Speaker Identification (SI) systems. Promising results have been obtained with Convolutional Neural Networks (CNNs). This success are mostly driven by the advent of large datasets. However in the context of commercial applications, collection of large amount of training data is not always possible. In addition, robustness of a SI system is adversely effected by short utterances. SI with only a few and short utterances is a challenging problem. Therefore, in this paper, we propose a novel text-independent speaker identification system. The proposed system can identify speakers by learning from only few training short utterances examples. To achieve this, we combine CNN with Scattering Wavelet Network. We propose a two-stage feature extraction framework using a two-layer wavelet scattering network coupled with a CNN for SI system. The proposed architecture takes variable length speech segments. To evaluate the effectiveness of the proposed approach, Timit and Librispeech datasets are used in the experiments. These conducted experiments show that our hybrid architecture performs successfully for SI, even with a small number and short duration of training samples. In comparaison with related methods, the obtained results shows that an hybrid architecture achieve better performance.**

I. Introduction

Speaker identification (SI) is an important biometric recognition technology. It is the task of identifying a person, based on a given speech signal and enrolled speaker records [1]. SI has gained great popularity in a wide range of applications, such as access user control, transaction authentication, forensics and personalization. After decades of research, significant performance improvement has been gained and some SI systems have been deployed in some practical applications [2], [3], [4]. In spite of these great achievements, current SI systems perform well only if the enrollment and test utterances are well matched, otherwise the performance will be seriously degraded. Moreover, many applications require very good accuracy even with short duration utterances. However, the performance of SI systems degrade with short utterances of about 5-10 seconds [5]. Different studies [6], [7] [8] have shown that the use of short segments may induce a drastic drop of the performances of authentication systems. This drop in performance is mainly due to the low amount of information on each speaker that is usually extracted from such short sequences. Speaker identification with only few and short utterances is thus a challenging problem.

Most of traditional SI systems are based on features relying on speech production and perception, such as Mel-Frequency Cepstral Coefficients (MFCCs), and on unsupervised generative models. During the training phase, MFCC features are used to train a Gaussian Mixture model (GMM) and to build an Universal Background Model (UBM) [9]. The GMM-UBM framework represents the speaker and channel independent attributes over their Gaussian components. However, it has been shown [10], [11] that it is beneficial to further process this vector by extracting intermediate vectors called i-vectors. During the authentication phase, an i-vector is extracted from a given speech sample and is compared to the reference i-vector, either with a simple cosine distance or with more complex techniques such as Probabilistic Linear Discriminant Analysis (PLDA) [12]. However, performance of these baseline methods suffer of sensitivity to lexical variability for short utterances [13].

Recently, deep learning has appeared in many pattern recognition fields. It has shown remarkable success in many fields such as image recognition [14] and natural language processing [15]. In speaker identification, a similar trend has been observed. Deep Neural Networks (DNNs) have been used with the i-vector framework to compute Baum-Welch statistics [16], or for frame-level feature extraction [17]. DNNs have also been proposed for direct discriminative speaker classification, as witnessed by the recent literature on this topic [18], [19]. Lately, there was an increasing number of studies trying the use of convolutional neural network [20] in numerous speech tasks [21], [22]. Some works have proposed to directly feed networks with spectrogram bins [23], [24] or even with raw waveforms [25], [26]. Among DNNs, CNNs have the most suitable architecture for processing raw speech samples, since weight sharing, local filters, and pooling constitute precious tools to discover robust and invariant representations. However, CNNs networks require numerous labeled training examples along with considerable computational resources and time to achieve effective learning. In a setting where only few labeled data with short duration are available, the training becomes difficult and requires a lot of regularization.

Recently, Mallat et al. [27] have proposed Scattering wavelet networks as a class of Convolutional Neural Networks (CNNs) with fixed weights. They have largely investigated the wavelet scattering transform (WST) framework and its properties. WST possesses the same properties used in CNN to extract reliable features from data. Additionally, the WST can extract reliable information at different scaling levels of decomposition. Also, it has been proved that the wavelet scattering coefficients are more informative than a Fourier transform when dealing with short variation signals or small deformation and rotation invariant [28], [29].

Scattering representations can be plugged into any classification or regression system, be it shallow or deep. The WST was tested on handwriting image data to extract the features where it achieved good performance [28]. WST has enjoyed significant success in various audio [27] and biomedical [30] signal classification tasks. WST demonstrated promising results on the TIMIT dataset for phonetic classification [31] and recognition [32].

In this paper, we propose a two-stage feature extraction framework using a two-layer wavelet scattering network coupled with a CNN for SI system. We explore the use of the WST for feature extraction along with a convolutional neural network. In this hybrid deep learning network, the use of a two-stage feature extraction framework can be helpful when there is a lack of data. This provides features of the same signal at different scales and captures its dominant energy. Such advantages could be useful when dealing with short duration utterances. The proposed network takes variable length speech segments. It is trained at the frame-level using the extracted features. The system has been evaluated on both Timit and Librispeech datasets and it has achieved better results than the state-of-the-art.

The remainder of this paper is organized as follows. Section II presents the wavelet scattering transform. Section III describes the proposed hybrid architecture, which is composed in a cascade of a scattering transform and a convolutional neural network. Section IV discusses the experimental setup and the corresponding results obtained by the proposed system as well as the ones provided by related systems.

II. Wavelet Scattering Transform

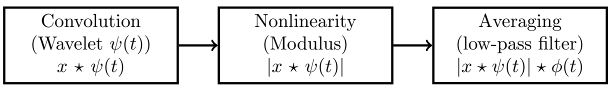

Fig. 1. Wavelet scattering transform processes, where x is the input data, ψ a wavelet function and ϕ an averaging low-pass filter.

To produce a wavelet scattering transform [27] of an input signal x, three successive main operations are required: convolution, nonlinearity, and averaging as described in Figure 1. The scattering transform coefficients are obtained with the averaging of wavelet modulus coefficients by a low-pass filter ϕ. Let a wavelet $\psi(t)$ be a band pass filter with a central frequency normalized to 1, and $\psi_\lambda(t)$ a wavelet filter bank, which is constructed by dilating the wavelet:

$$\psi_\lambda(t) = \lambda\psi(\lambda t) \tag{1}$$

where $\lambda = 2^{\frac{j}{Q}}$, $\forall j \in Z$ and Q is the number of wavelets per octave.

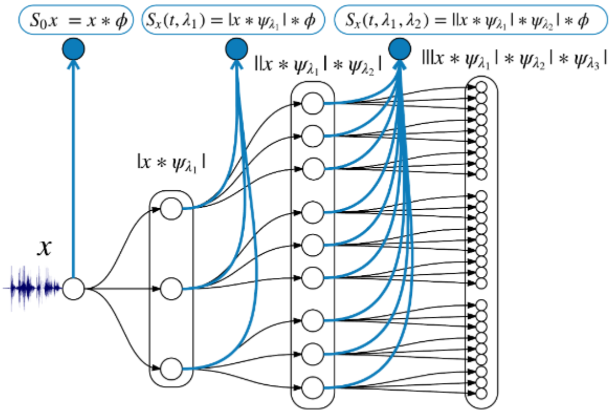

Fig. 2. Hierarchical representation of scattering coefficients at multiple layers [27].

The bandwidth of the wavelet $\psi(t)$ is of the order $\frac{1}{Q}$, and as a result, the filter bank is composed of band pass filters which are centered in the frequency domain in λ and have a frequency bandwidth $\frac{\lambda}{Q}$.

At the zero order, we have a single coefficient given by $S_0x(t) = x \star \phi(t)$, which is close to zero for audio signals. At the first order, we have:

$$S_1x(t, \lambda_1) = |x \star \psi_{\lambda_1}| \star \phi(t) \tag{2}$$

The second order coefficients capture the high-frequency amplitude modulations occurring at each frequency band of the first layer and are obtained by:

$$S_2x(t, \lambda_1, \lambda_2) = ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi(t) \tag{3}$$

Figure 2 shows the hierarchy of scattering coefficients. This somewhat resembles to the structure of deep neural networks, although that in the scattering transform, each layer provides some output, while the only output of most of deep neural networks is provided by the last layer. This decomposition on first and second orders scattering coefficients is applied to the time domain signals. Second order features are normalized by first order features, to ensure that the higher order of scattering depends on the amplitude modulation component of the speech signal. The first and second orders of the scattering transform are concatenated to form a scattering feature vector for a given frame. The scattering features include log-mel features al together with higher order features to preserve the greater details in the speech signal [27]. This representation is invariant to time shifts and is stable to deformations. Hence, to ensure invariability to frequency translation on a logarithmic scale like translation of speaker formants, a scattering transform is performed along log-frequency. The logarithm is applied to each coefficients of the scattering feature vector. It is thus locally translation invariant in time and log frequency, and stable to time and frequency deformations.

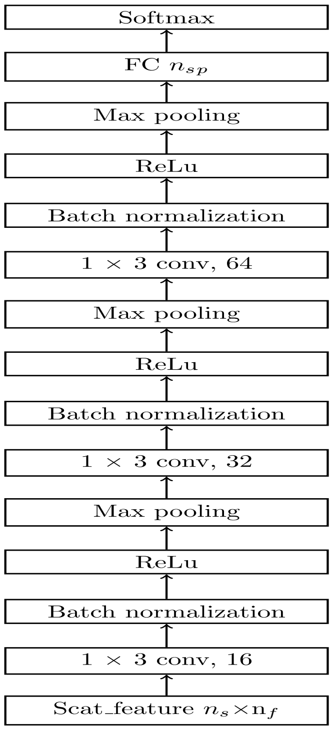

Fig. 3. The proposed Hybrid Wavelet Scattering Transform Convolutional Neural Network (HWSTCNN) architecture. It consists of two parts: feature extraction, where a scattering network is coupled with a convolutional layers to extract frame level features, and dense classification layer.

III. Hybrid Network Architecture

An ideal model for SI system should take variable length speech segments and produce a discriminating output descriptor. The distance between descriptors of different speakers must be larger than those of the same speaker. To satisfy all mentioned properties, Figure 3 shows the proposed Hybrid Wavelet Scattering Transform Convolutional Neural Network (HWSTCNN). The network consists of two parts: feature extraction and frame level embedding. The scattering network is coupled with convolutional layers to extract frame level features, and dense classification layer to construct speaker frame embedding.

The network is shown in Figure 3 and described in more details in the following paragraphs. The proposed architecture consists in two scattering network layers, namely, Scat-Layer, three 1D convolutional layers, and in one fully connected layer. Scat-Layer performs scattering wavelet transform on overlapping frames (500ms with 125ms skip rate) in time-domain signal. After the Scat-Layer, three convolutional layers are followed by one fully connected layer. Standard CNN pipeline (pooling, batch normalization, ReLU activation) was employed. Final softmax layer performs speaker classification.

Scat-Layer is composed of two scattering wavelet transform layers. The first layer contains 8 Gabor wavelets per octave and the second has one Morlet wavelet per octave. This configuration was chosen to match the frequency resolution of

TABLE I
HWSTCNN architecture. Each row specifies the # of convolutional filters, their sizes, and the # filters.

Layer name	**Hybrid model**	**Output**
Input	–	$n \times 1$
ScatNet layer	scat at 2	$n_s \times n_f \times 1$
Conv1 block	conv1D, 3×1,16 bn relu	$n_s \times n_f \times 16$
Pooling	maxpool, 2×1, stride (2,1)	$n_s/2 \times n_f/2 \times 16$
Conv2 block	conv1D, 3×1,32 bn relu	$n_s/2 \times n_f/2 \times 32$
Pooling	maxpool, 2×1, stride (2,1)	$n_s/4 \times n_f/4 \times 32$
Conv3 block	conv1D, 3×1,64 bn relu	$n_s/4 \times n_f/4 \times 64$
Pooling	maxpool, 2×1, stride (2,1)	$n_s/8 \times n_f/8 \times 64$
Embedding	fc,n_{sp}	n_{sp}
Loss	softmax	

Mel filters at the first level. The second order of the scattering transform recovers the lost information. Averaging window length was set to 64ms. Later, coefficients are normalized and log-transformed. Therefore, the representation of speech signal using the first and the second orders of the scattering transform extends the MFCC representation and doesn't loose information. These scattering coefficients are computed a publicly available toolbox [27].

Each convolutional layer is formed by a 1D filter of length 3 and batch normalization. They are followed by a max-pooling layer, with pooling size 1×2 and stride 1×2. The number of filters is respectively 16, 32 and 64. A fully connected layer with n_{sp} hidden neurons, where n_{sp} is the number of speakers to be identified, is connected to categoral softmax layer. The softmax produces a probability distribution per frame over the target speakers in the dataset.

We use rectified linear units as activation functions in all layers. Stochastic gradient descent was used as an optimizer with a learning rate of 0.001 and 0.9 momentum. The network is trained with mini batches of size 64 for 10 epochs. The proposed architecture is shown in Figure 3 and details such as the number of filters and kernel sizes are summarized in Table I. This architecture takes raw speech frames with time-windows of 500ms and with a skip rate of 125ms to produce speaker embedding at frame-level. The amount of parameters in this neural network is 18,1 millions which is less than the actual state-of-the-art, as it will be shown in the next sections. Coupling a scattering network with a convolutional network for building our hybrid architecture can reduce the instabilities in the first layers as the wavelet scattering transform is stable and non-expansive. By reducing the variability at feature extraction stage, the proposed hybrid architecture can generate discriminative feature information at frame level. This hybrid architecture has the capability to reduce the required depth and spatial dimension of the deep learning networks, which makes the strength of using both scattering transform and CNN.

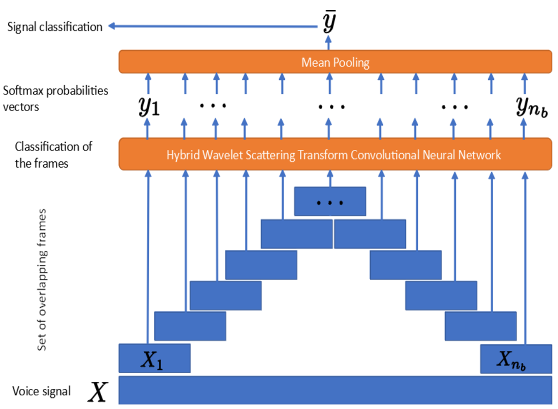

Fig. 4. Hybrid Wavelet Scattering Transform Convolutional Neural Network.

In the testing phase of our system, a speech utterance X to be classified is divided into m overlapping frames of length n_f. Each frame shares the first part with the previous frame and the last part with the next frame. Each frame x_i is fed to our HWSTCNN to predict the frame speaker label. Softmax probabilities $Pr(s_k|x_i)$ are calculated to estimate if the frame x_i is from the speaker s_k among the n_{sp} speakers. These speaker membership probabilities $y_i \in R^{n_{sp}}$ constitute a vector for each frame. Finally, the mean membership probability vector $\bar{y}$ of the whole utterances is given by the mean of all the stored probability vectors computed per frame. The estimated speaker label for the whole speech utterance X corresponds to the speaker of maximum of probability: $label = \arg\max_{j\in[1,n_{sp}]} \bar{y}^j$ with $\bar{y}^j$ the j^{th} column of the $\bar{y}$ vector and n_{sp} the total number of speakers.

Figure 4 describes the process of labeling speech utterances. Speech frames are first fed into our HWSTCNN to predict frame speaker probability vectors. These frame-level predictions are then aggregated into a whole-utterance-level prediction using the mean of the obtained softmax frame-level probabilities. Therefore the speaker's probabilities prediction is initially made per frames and thereafter is converted into a final single speaker vector probability for the whole utterance. Finally, the estimated speaker label corresponds to the one that has the highest probability among the ones of the average vector probability $\bar{y}$.

IV. Experiments

This section describes the experiments and the results obtained with our approach and related systems.

A. Dataset and experimental setting

Two datasets are used in the experiments, TIMIT [34] and LibriSpeech [35].

- The TIMIT dataset contains studio quality recordings of 630 speakers (192 female, 438 male), sampled at 16 kHz, covering the eight major dialects of American English. Each speaker reads ten phonetically rich sentences. We consider only 462 speakers from TIMIT. We use only 8 sentences for each speaker, the “SX” (5 sentences) and the “SI” (3 sentences). The “SX” sentences are used to train the system, while the “SI” sentences are used to test.
- The LibriSpeech database consists in audio books read-out-loud by 2484 speakers, 1283 male and 1201 female volunteers who recorded their voices spontaneously. The speech signal is usually clean, but the recording device and channel conditions vary a lot between different utterances and speakers. We decided to keep 7 utterances of each speaker for training, and 3 utterances as a fixed test set for evaluation.

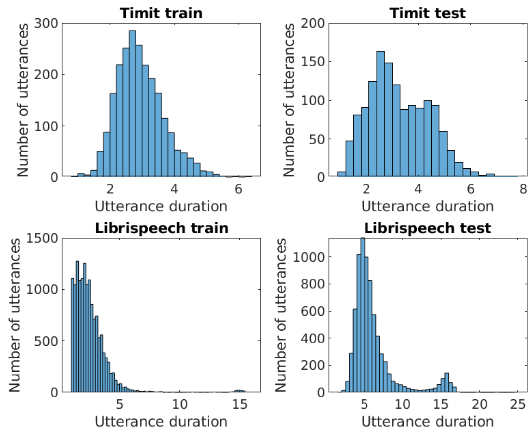

Fig. 5. Distribution of the utterance lengths in Timit and Librispeech databases.

TABLE II

Number of utterances for the speaker identification task.

	Train	Test	Total
Timit	2310	1386	3696
Librispeech	14904	7452	22356

In Timit and Librispeech datasets, the duration of training sentences is about 12-15 seconds for each speaker and test sentences duration is about of 2-5 seconds. Figure 5 shows the distribution of the utterances durations. The average duration was of 4s, the minimum was of 1s, and the maximum was of 17s. Utterances with durations of less than 4 seconds represented about 87% of the data. Table II presents the number of training and testing utterances for both datasets.

To validate the effectiveness of our model, we built 8 kHz and 16 kHz versions of our system. Timit and Librispeech datasets are downsampled to 8 kHz. Finally, we evaluate the performance of the proposed system under the noisy environment. To that aim, we have made noise-corrupted versions of the Timit training database by adding different types of noise at different SNR level. The noisy utterance for training were made by adding white or babble noises of the SNR levels of 5dB, 10dB and 20dB. However, the noisy utterances for the test were obtained by one of two mentioned noise types of the SNR levels of 5dB, 10dB and 20dB respectively.

Experiments are conducted on the full and short length conditions. We did not apply any pre-processing to the raw waveforms, such as pre-emphasis, silence removal, detection and removal of unvoiced speech. Scattering transform was computed to the depth of 2 with speech frames of 64ms of length. The first layer contained 8 Gabor wavelets per octave and the second had one Morlet wavelet per octave. The averaging window was set to 500ms of length. Later, coefficients are normalized and log-transformed. Stochastic gradient descent was used as an optimizer with a learning rate of 0.001 and 0.9 momentum. The network is trained with mini batches of size 64 for 10 epochs. Our implementation is based on Scatnet [27] and deep learning Matlab toolboxes.

B. Related systems

In order to evaluate the performance of our proposed system, two alternative state-of-the-art systems were investigated: SincNet [36] and CNN-Raw [37] systems for speaker identification.

- Sincnet is a novel end-to-end neural network architecture, that directly receives raw waveforms as input. The first 1D convolutional layer of SincNet is composed by Sinc functions. SincNet convolves the waveform with a set of parametric sinc functions that implement band-pass filters. The filters are initialized using the Mel-frequency filter bank and their low and high cutoff frequencies are adapted with standard back-propagation as any other layer. The first layer performs Sinc based convolutions, using 80 filters of length 251. The remaining two layers use 60 filters of length 5. Next, three fully-connected layers composed of 2048 neurons and normalized with batch normalization are applied. All hidden layers use leaky-ReLU non-linearity. Frame-level binary classification is performed by applying a softmax classifier and cross-entropy criteria [36]. The number of parameters in Sincnet is about 26,5 millions. For training, this architecture needs about 2900 epochs to converge on the Librispeech dataset.
- In the CNN-Raw system, the raw waveform is fed directly to the first layer. Three convolution layers are used to perform the feature mapping. Each convolution layer is composed of 80 filters followed by a max pooling. Next, three fully-connected layers composed of 2048 neurons and normalized with batch normalization are applied. All hidden layers use leaky-ReLU non-linearities. Frame-level binary classification is performed by applying a softmax classifier and cross-entropy criteria [37]. Both networks are trained with 800 epochs and batches of size 128. The number of parameters in CNN-raw is about 27,6 millions. The number of parameters in CNN-raw and Sincnet is larger compared to our proposed architecture.

Table III summarizes the number of learning parameters of all tested methods. We observe that the number of learning parameters required by our method is lower than the ones of SincNet and CNN-Raw by about 33%.

TABLE III
NUMBER OF PARAMETERS AND EPOCHS FOR OUR SYSTEM AND RELATED SYSTEMS.

	SincNet	**CNN**	**Proposed**
Parameters $\times 10^6$	26,5	27,6	**18,1**

TABLE IV
IDENTIFICATION ACCURACY RATE (%) OF THE PROPOSED SPEAKER IDENTIFICATION ON 8K AND 16K DATA TRAINED AND TESTED WITH FULL UTTERANCES.

	8k	**16k**
LibriSpeech	97.38	99.28
TIMIT	85.93	98.12

C. Results

In order to evaluate our proposed speaker identification system we use the identification accuracy rate which is equal to the number of correct identifications over the number of speakers to test. In Table IV, we report the effect of sampling frequency on system performance. As expected, results show that our system performs better on 16 kHz than 8 kHz data. However, correct identification rates with 8KHz data are smaller than rates with 16KHz by only 2%. Our system remains thus competitive for low sampling frequency rate.

TABLE V
IDENTIFICATION ACCURACY RATE (%) OF THE PROPOSED SPEAKER IDENTIFICATION AND RELATED SYSTEMS TRAINED AND TESTED WITH FULL UTTERANCES.

	LibriSpeech	**TIMIT**
CNN-raw	69.82	60.53
SincNet-raw	79.32	61.81
Proposed	**99.28**	**98.12**

The correct identification rates for different methods are shown in Table V. Results are compared on both TIMIT and Librispeech datasets. Results from this table shows that our hybrid network outperforms both SincNet and CNN-Raw systems. We observe that our methods obtains significant robustness on both datasets. On TIMIT, our system achieves a relative improvement of about 38% over CNN-raw and 36% over SincNet. Moreover on LibriSpeech, our system achieves a relative improvement of about 30% over CNN-raw and 20% over SincNet. The table shows that the effect of coupling the first convolution layer of CNN with WST, improves the identification performance.

We further investigate the performances of our system in Table VI. We use the same deep CNN architecture used in Sincnet. An increase of 19% is proved in accuracy performance. This shows the benefit of using the scattering transform.

We report in Table VII the effect of training utterances duration per speaker on performances. We split the training

TABLE VI
IDENTIFICATION ACCURACY RATE (%) OF THE PROPOSED SPEAKER IDENTIFICATION AND RELATED SYSTEMS TRAINED WITH SAME MODEL ON LIBRISPEECH.

	SincNet-raw	CNN-raw	Proposed
full-full	79.32	69.82	**99.46**

TABLE VII
IDENTIFICATION ACCURACY RATE (%) OF THE PROPOSED SPEAKER IDENTIFICATION ON LIBRISPEECH 16KHZ DATASET TRAINED AND TESTED WITH DIFFERENT UTTERANCES DURATIONS.

	Train utterance duration		
Test	**8s**	**12s**	**full**
1.5s	96.86	97.20	97.38
3s	98.76	98.93	98.97
full	99.12	99.25	99.28

data to obtain a total duration of 8s or 12s per speaker. Full train duration is about 14s. This table depicts the correct identification rates for 1.5s, 3s and full duration of testing utterances. We observe that the proposed methods obtain significant robustness, which indicates that the proposed method is able to extract speaker identity features in different training and testing conditions. Our system gives higher accuracy rate for all conditions of short-utterance task. As shown in Table VII, varying the number of samples per speaker and thus the total duration for training induces a variation of only 0.15% of the accuracy. On the other hand, using 3s duration instead of 1.5s induces an small increase of the accuracy of about 0.5%. Our system is thus able to construct discriminating speakers models with few number of training data but provides better results with test and train samples of at least 1.5s.

TABLE VIII
IDENTIFICATION ACCURACY RATE (%) OF THE PROPOSED SPEAKER IDENTIFICATION TRAINED AND TESTED IN NOISY CONDITIONS.

SNR	White	Babble
20	68.54	89.03
10	45.67	84.70
5	33.62	77.63

In Table VIII, we report noise effect on our SI system. The results show that our proposed method achieves acceptable performance across all the SNR levels on Timit database. Identification with speech emerged on babble noise gives good accuracy.

V. CONCLUSION

In this paper, we have proposed a speaker identification system that learns speaker discriminating information directly from raw speech signals using scattering network and CNNs. We have explored the potential advantage of WST in extracting robust speaker representation. We have demonstrated that by coupling CNN with scattering network, we are able to compute a stable description of speaker identity information. Experimental results on TIMIT and Librispeech corpuses have shown that the proposed method can achieve dominant results in clean condition. We have shown the effectiveness of our hybrid architecture for speaker identification with different utterances duration. Our results show that our hybrid method yields significant improvements over SincNet and CNN-Raw methods on the same databases. This work could be extended in future works by applying the attention mechanism to generate utterance level embedding.

REFERENCES

[1] H. Beigi, "Fundamentals of Speaker Recognition," Springer US, 2011.
[2] D. A. Reynolds, "An overview of automatic speaker recognition technology," in Acoustics, speech, and signal processing (ICASSP), 2002 IEEE international conference on, vol. 4. IEEE, 2002, pp. IV–4072.
[3] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," Speech communication, vol. 52, no. 1, pp. 12–40, 2010.
[4] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," IEEE Signal processing magazine, vol. 32, no. 6, pp. 74–99, 2015.
[5] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "Plda for speaker verification with utterances of arbitrary duration," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, May 2013, pp. 7649–7653.
[6] A. Poddar, M. Sahidullah, and G. Saha, "Speaker verification with short utterances: a review of challenges, trends and opportunities," IET Biometrics, vol. 7, no. 3, pp. 91–101, 2018.
[7] Rohan Kumar Das and S. R. M. Prasanna, "Speaker verification for variable duration segments and the effect of session variability," Lecture Notes in Electrical Engineering, pp. 193–200, 2015.
[8] A. Kanagasundaram, R. Vogt, D. Dean, and S. Sridharan, "i-vector based speaker recognition on short utterances," in Proc. of Interspeech , 2011.
[9] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," Digital Signal Processing, vol. 10, no. 1, pp. 19–41, 2000.
[10] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouel- let, "Front-end factor analysis for speaker verification," IEEE Transactions on Audio, Speech, and Language Processing , vol. 19, no. 4, pp. 788–798, 2011.
[11] Wei Li, Tianfan Fu, and Jie Zhu, "An improved i-vector ex- traction for speaker verification," EURASIP Journal on Audio, Speech and Music Processing , pp. 1–9, 2015.
[12] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in Proc. of International Conference on Computer Vision , 2007.
[13] W. Li, T. Fu, H. You, J. Zhu, and N. Chen, "Feature sparsity analysis for i-vector based speaker verification," Speech Communication, vol. 80, pp. 60–70, 2016.
[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Proc. of Advances in Neural Information Processing Systems , 2012.
[15] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in Proc. of the International Conference on Machine Learning , 2008.
[16] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, "Deep neural networks for extracting baum-welch statistics for speaker recognition," in Proc. of Speaker Odyssey , 2014.
[17] S. Yaman, J. W. Pelecanos, and R. Sarikaya, "Bottleneck features for speaker recognition," in Proc. of Speaker Odyssey, 2012, pp. 105–108.
[18] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small foot-print text-dependent speaker verification," in Proc. Of ICASSP , 2014, pp. 4052–4056
[19] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in Proc. of ICASSP , 2018.
[20] Yann LeCun and Yoshua Bengio, "Convolutional networks for images, speech, and time series," in The hand-book of brain theory and neural networks , 1995, vol. 3361, p. 1995.
[21] Abdel-Hamid O., Mohamed Abdel-rahman, Jiang Hui, Deng Li, Penn Gerald, and Dong Yu, "Convolutional neural networks for speech recognition," IEEE Transactions on Audio, Signal, and Language Processing , vol. 22, no. 10, pp. 1533–1545, 2014.

[22] T. N. Sainath, A. Mohamed, B. Kingsbury, and B. Ramabhad-ran, "Deep convolutional neural networks for LVCSR," in Proc. of ICASSP , 2013.
[23] C. Zhang, K. Koishida, and J. Hansen, "Text-independent speaker verification based on triplet convolutional neural network embedding," IEEE Transactions on Audio, Signal, and Language Processing , vol. 26, no. 9, pp. 1633–1644, 2018.
[24] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb:a large scale speaker identification dataset," in Proc. of Interspeech, 2017.
[25] D. Palaz, M. Magimai-Doss, and R. Collobert, "Analysis of CNN-based speech recognition system using raw speech as input," in Proc. of Interspeech , 2015.
[26] H. Muckenhirn, M. Magimai-Doss, and S. Marcel, "To-wards directly modeling raw speech signal for speaker verification using CNNs," in Proc. of ICASSP , 2018.
[27] Joakim Andèn and Stephane Mallat, "Deep scattering spectrum," IEEE Transactions on Signal Processing , vol. 62, no. 16, pp. 4114–4128, 2014.
[28] E. Oyallon and S. Mallat, "Deep roto-translation scattering for object classification," in Proc. of IEEE Conference on Computer Vision and Pattern Recognition , 2015, pp. 2865–2873.
[29] Stephane Mallat, "Group invariant scattering," Communications on Pure and Applied Mathematics , vol. 65, no. 10, pp. 1331–1398, 2012.
[30] Vaclav Chuda cek, Joakim And en, Stephane Mallat, Patrice Abry, and Muriel Doret, "Scattering transform for intra-partum fetal heart rate variability fractal analysis: A case-control study," IEEE Transactions on Biomedical Engineering vol. 61, no. 4, pp. 1100–1108, 2014.
[31] V. Peddinti, T. N. Sainath, S. Maymon, B. Ramabhadran, D. Nahamoo, and V. Goel, "Deep scattering spectrum with deep neural network," in Proc. of ICASSP , 2014, pp. 361–364.
[32] P. Fousek, P. Dognin, and V. Goel, "Evaluating deep scattering spectra with deep neural networks on large-scale spontaneous speech task," in Proc. of ICASSP , 2015, p. 54.
[33] M. Oquab, L. Bottou, I. Laptev, J. Sivic, "Weakly supervised object recognition with convolutional neural networks," In NIPS, 2014.
[34] L. Lamel, and R. Kassel, and S. Seneff, "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," Proc. of DARPA Speech Recognition Work-shop, 1986.
[35] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," Proc. of ICASSP, pp. 5206–5210, 2015.
[36] M. Ravanelli and Y. Bengio, "Speaker Recognition from raw waveform with SincNet," Proc. of SLT, 2018.
[37] H. Muckenhirn, M. Magimai-Doss, and S. Marcel, "On Learning Vocal Tract System Related Speaker Discriminative Information from Raw Signal Using CNNs," Proc. of Interspeech, 2018.