Received: 14 November 2017 | Revised: 5 February 2018 | Accepted: 10 April 2018

DOI: 10.4218/etrij.2017-0260

ORIGINAL ARTICLE

WILEY ETRI Journal

Dysarthric speaker identification with different degrees of dysarthria severity using deep belief networks

Aref Farhadipour[1] | **Hadi Veisi**[2] | **Mohammad Asgari**[1] | **Mohammad Ali Keyvanrad**[3]

[1]Department of Media Engineering, IRI Broadcast University, Tehran, Iran.

[2]Faculty of New Sciences and Technologies, University of Tehran, Tehran, Iran.

[3]Department of Computer Engineering and Information Technology, Amirkabir University of Technology, Tehran, Iran.

Correspondence
Aref Farhadipour, Department of Media Engineering, IRI Broadcast University, Tehran, Iran.
Email: areffarhadi@gmail.com

Dysarthria is a degenerative disorder of the central nervous system that affects the control of articulation and pitch; therefore, it affects the uniqueness of sound produced by the speaker. Hence, dysarthric speaker recognition is a challenging task. In this paper, a feature-extraction method based on deep belief networks is presented for the task of identifying a speaker suffering from dysarthria. The effectiveness of the proposed method is demonstrated and compared with well-known Mel-frequency cepstral coefficient features. For classification purposes, the use of a multi-layer perceptron neural network is proposed with two structures. Our evaluations using the universal access speech database produced promising results and outperformed other baseline methods. In addition, speaker identification under both text-dependent and text-independent conditions are explored. The highest accuracy achieved using the proposed system is 97.3%.

KEYWORDS
deep belief network, deep neural network, dysarthria, MFCC, speaker identification

1 | INTRODUCTION

Dysarthria is a motor speech disorder that affects the control of speech organs, which in turn affects different levels of speech production, such as the intelligibility of speech, control of articulation and pitch, phonation, and language production. It is associated with unusual phonation and amplitude, especially on explosive phonemes [1]. Dysarthria may be congenital or may be triggered by other diseases, causing unnatural variations in the voices of dysarthric individuals. The human auditory system may not be able to successfully identify these people by their voices because of these unnatural variations. Because dysarthric individuals are often physically incapacitated and unable to use a computer, speech processing applications, such as speech recognition and speaker recognition systems, can be helpful for dysarthric individuals. Although the accuracy of novel speaker-recognition systems for normal speakers is acceptable [2], the performance of these systems declines significantly for individuals suffering from dysarthria. Because of this unsatisfactory performance in terms of accuracy, it is necessary to design a new speaker-recognition system for dysarthric individuals. However, speaker recognition for dysarthric individuals is a challenging issue [3,4]. To the best of our knowledge, only one significant study has been reported in the literature on this issue [5], where a hybrid Gaussian mixture model-support vector machine (GMM-SVM) system was proposed for dysarthric speaker identification. In this work, the two dysarthria databases of Nemours [6] and TORGO [7] were used to train and evaluate the proposed system. These two databases have different characteristics with respect to speaker age and environmental conditions. The best result achieved in this study was a speaker identification rate of 97.2% [5]. Our work is the first attempt at dysarthric speaker identification with different intelligibility levels using neural

networks on the universal access (UA) database [8]. Contrary to dysarthric speaker recognition, there have been several studies on dysarthric speech recognition [9–11].

In speaker identification, the task is to identify an unknown speaker from a set of known speakers. However, owing to the high inter-speaker variability for dysarthric individuals, these systems require robust signal representation and modeling. Deep belief networks (DBNs) are an effective technique for robust feature extraction and modeling [12], and have recently been successfully implemented in speech processing applications [13]. Deep belief network is a powerful hierarchical generative model that can be applied to the acoustic modeling of speech signals.

In this work, we propose the use of a DBN auto-encoder to realize the feature extraction of the speech signal of dysarthric speakers. To this end, Mel frequency cepstral coefficients (MFCCs) were encoded using a DBN, and they were then identified using a multi-layer perceptron (MLP).

The rest of the paper is organized as follows: In Section 2, the theoretical background related to the paper is explained. Here, feature extraction, DBNs, and speaker modeling are reviewed. In Section 3, the proposed method for dysarthric speaker identification is presented. Then, in Section 4, the experimental settings and the results are given. Finally, Section 5 concludes this paper.

2 | THEORETICAL BACKGROUND

2.1 | Dysarthric speaker

Speech is a complex process requiring the synchronous and timely contraction of a large number of muscle groups associated with respiration, laryngeal function, and articulation. Disorders that affect each of these components result in the debilitation of speech and inherent speech features of individual speakers. Dysarthria is a motor speech disorder, and is defined as a collective term to represent a group of related speech disorders that are caused by disturbances in the muscular control of the speech mechanism [1]. It results from the impairment of any of the basic motor processes involved in speech production, affecting respiration, phonation, resonance, articulation, and prosody [14]. Speech disorders significantly affect one's lifestyle by inhibiting individuals from expressing opinions and needs; they also affect personality as well as self-esteem and relationships. Speech impairments can result from damage to the central or peripheral nervous systems, leading to weak and slow muscles, incoordination of muscular movements, altered muscle tone, and inaccuracy of oral and vocal movements [15]. This may result in abnormal characteristics of speech quality and reduced intelligibility of speech. The poor quality of speech produced by a dysarthric speaker makes it difficult for automatic processing and identification.

2.2 | Speaker identification

Human beings can reliably recognize known voices by hearing only a few seconds of speech. The uniqueness of an individual's voice can be attributed to both the physical and acquired characteristics of a person. Physical differences exist largely owing to the distinct shapes and sizes of the voice-producing organs (for example, the vocal tract and larynx) and partly owing to the articulators (for example, the tongue, teeth, and lips). Apart from these anatomical properties, individuals can also be distinguished by their accent, vocabulary, speech rate, and other personal mannerisms that are acquired over a period of time in their speech patterns. State-of-the-art speaker-recognition systems exploit these properties in order to achieve a high recognition accuracy [16]. Speaker recognition (SR) can be broadly divided into two categories, namely, speaker identification (SI) and speaker verification (SV). The task of validating the claimed identity of a speaker is known as SV. The task of detecting a unique speaker responsible for producing a given utterance, out of a closed set of enrolled speakers is known as SI. The main elements of an SI system are shown in Figure 1. Considering each speaker model as a class, the SI task is basically a multi-class classification problem in which an unknown test utterance is assigned to a particular class. In an SI system, the training phase tries to estimate the acoustic models of individual speakers, and the test phase performs the identification of an unknown speaker using the trained acoustic models.

An SI system may be text-dependent (TD) or text-independent (TI). In TD systems, the recognition words are fixed or known in advance; however, in the TI case, there are no constraints on the phrases that the speakers are allowed to speak. Therefore, in TI systems, the train and test utterances may have completely different contents, and the recognition systems should therefore consider this phonetic mismatch. In the next subsections, the main steps of an SI system are briefly described.

2.2.1 | Preprocessing

This stage corresponds to the acquisition of a speech signal from the microphone and the digitization of the analog speech signal. For several SI tasks, a “voiced activity detector” (VAD) module is often used [17]. The VAD module attempts to separate speech signals from silence and background noises, such that only the speech segments must be processed. The VAD is an important front-end of modern speaker-recognition systems that can reduce the computational load and the probability of false classifications. Moreover, the VAD can shift the detected boundaries of the beginning and end of an utterance to guarantee that the entire phrase is enclosed for pattern recognition. In this

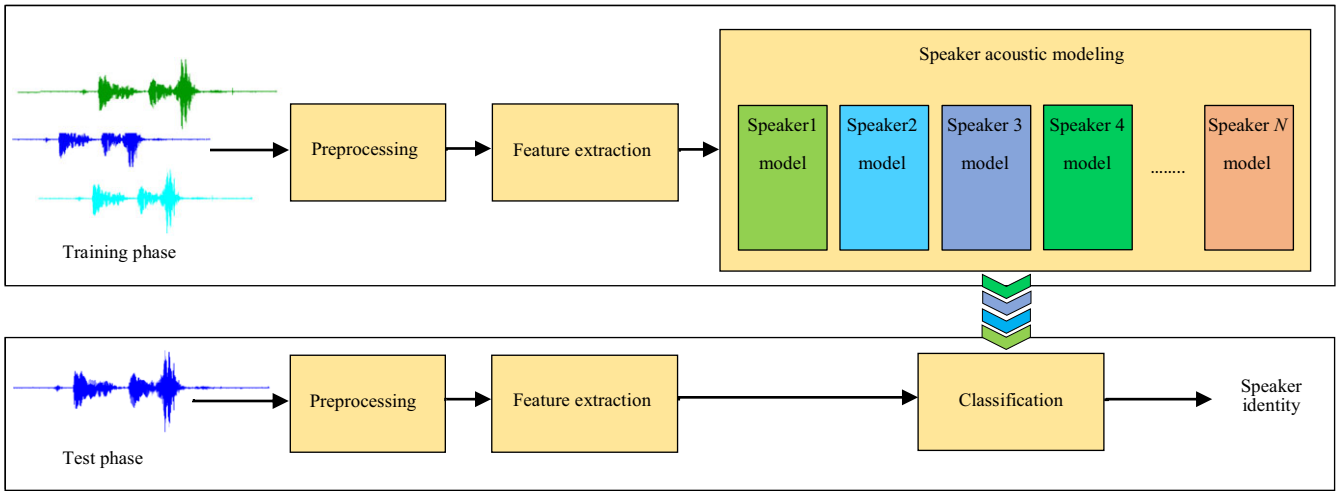

FIGURE 1 General components of speaker identification system

research, we employ a statistical model-based VAD proposed by Sohn and others [18].

2.2.2 | Feature extraction

In this step, a set of parameters is extracted from a raw speech signal. Ideal features for speaker identification have large intra-speaker variability and small inter-speaker variability. These features occur frequently and naturally in speech, and are not affected by a speaker's health or long-term variations in voice. They are channel and context independent, and are easily extracted from the speech signal. Apart from these, the features should have a compact representation in order to avoid using a large amount of training data. The selection of appropriate features for SR is usually based on certain criteria, such as the intended application, computing resources, and the amount of speech data available. Short-term spectral features such as MFCCs [19] and perceptual linear prediction (PLP) [20] are often preferred for SR tasks because they are easy to compute, and demonstrate good performance under clean and controlled acoustic conditions. However, these features are susceptible to noise as well as channel and acoustic variations. Therefore, the development of robust features remains a problem.

In this work, we have proposed a DBN auto-encoder for feature extraction. The effectiveness and robustness of these features have been proven in previously developed speech processing applications such as speech recognition [12]. A DBN consists of a stacked restricted Boltzmann machine (RBM). An RBM consists of a layer of stochastic binary visible units connected to a layer of stochastic binary hidden units that learn to model distribution over visible units [12]. In this machine, there are connections between the visible and hidden units, but there are no visible-visible or hidden-hidden connections, as shown in Figure 2. The learning algorithm of an RBM is reviewed below.

A joint configuration, (v, h) of the visible and hidden binary units of an RBM has an energy given by

$$E(v,h) = -\sum_{i=1}^{V}\sum_{j=1}^{H} w_{ij}v_i h_j - \sum_{i=1}^{V} a_i v_i - \sum_{j=1}^{H} b_j h_j, \quad (1)$$

where v_i, h_j are the binary states of visible unit i and hidden unit j, respectively, w_{ij} represents the symmetric interaction between v_i and h_j, which are model parameters, and a_i and b_j are bias terms. Depending on their energy, each configuration is associated with a probability given as

$$P(v,h|\lambda) = \frac{e^{-E(v,h)}}{Z} \quad (2)$$

where $Z = \sum_v \sum_h e^{-E(v,h)}$ is called the partition function, and performs normalization, and λ denotes the set of model parameters. Therefore, we can write an expression for the marginal probability by assigning an RBM to some visible vector v,

$$P(v|h) = \frac{\sum_h e^{-E(v,h)}}{Z}. \quad (3)$$

The derivative of the log probability $P(v|\lambda)$with respect to the model parameters, w_{ij}, is

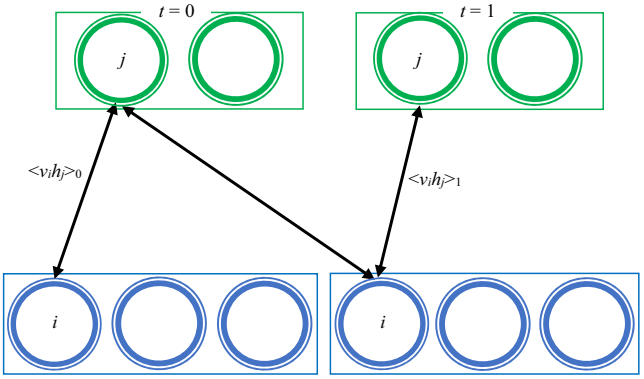

FIGURE 2 One step contrastive divergence of an restricted Boltzmann machine

$$\frac{\partial \log P(v|\lambda)}{\partial w_{ij}} = \langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{model}} \quad (4)$$

where $\langle \alpha \rangle_{\text{data}}$ and $\langle \alpha \rangle_{\text{model}}$ are the expectation of α estimated from the data and model, respectively. The derivative in (4) leads to the following learning rule:

$$\Delta w_{ij} = \varepsilon \left(\langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{model}} \right) \quad (5)$$

where ε is the learning rate.

The hidden neurons are conditionally independent given the visible vector. Then, the binary state of each hidden unit h_j is set to one with probability

$$P(h_j = 1|v) = \psi \left(\sum_i w_{ij} v_i + b_j \right) \quad (6)$$

where $\psi(x) = \frac{1}{1+e^{-x}}$ is the sigmoid logistic function. For the binary visible neuron, a completely symmetric derivation allows us to obtain

$$P(v_i = 1|h) = \psi \left(\sum_j w_{ij} h_j + a_i \right). \quad (7)$$

The estimation of $\langle v_i h_j \rangle_{\text{data}}$ based on (6) and (7) is straightforward because of the absence of direct connections between the same units in an RBM, but the exact calculation of the $\langle v_i h_j \rangle_{\text{model}}$ term takes exponentially long time when the hidden values are unknown. Therefore, to estimate this term, the use of approximated methods, such as contrastive divergence (CD) [21], is required.

The one-step CD of an RBM is shown in Figure 2. The approximation for the gradient with regard to the visible hidden weights is

$$\begin{aligned} \Delta w_{ij} &= -\varepsilon \left(\langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\infty} \right) \\ &\approx -\varepsilon \left(\langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{1} \right). \end{aligned} \quad (8)$$

where $\langle . \rangle_{\infty}$ denotes the expectation computed with samples that are generated by running the Gibbs sampler in infinite steps, and $\langle . \rangle_1$ denotes the expectation for running in one step. Similarly, the learning rules for bias parameters are

$$\begin{aligned} \Delta a &= -\varepsilon \left(\langle v \rangle_{\text{data}} - \langle v \rangle_{1} \right) \\ \Delta b &= -\varepsilon \left(\langle h \rangle_{\text{data}} - \langle h \rangle_{1} \right). \end{aligned} \quad (9)$$

RBMs can also be applied to model the distribution of real-valued data such as MFCC vectors. When the visible units v are real-valued and hidden units h are binary, the RBM energy function can be modified to enable it to adapt to such variables, giving a Gaussian–Bernoulli RBM; the energy is defined as [12]

$$E(v,h) = \sum_{i=1}^{V} \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_{i=1}^{V} \sum_{j=1}^{H} \frac{v_i}{\sigma_i} w_{ij} h_j - \sum_{j=1}^{H} b_j h_j \quad (10)$$

where the variance parameters σ^2 are commonly fixed to a predetermined value instead of being learned from training data. In order to train a Gaussian–Bernoulli RBM using the CD algorithm, two conditional distributions for Gibbs sampling are derived as follows.

$$P(h_j = 1|v) = \psi \left(b_j + \sum_i \frac{v_i}{\sigma_i} w_{ij} \right), \quad (11)$$

$$P(v_i|h) = N \left(v, \sum_j h_j w_{ij} + a_i, \sigma_i^2 \right) \quad (12)$$

where $N(v, \mu, \Sigma)$ denotes a Gaussian distribution of v with a mean vector μ and a covariance matrix Σ. For unsupervised pre-training using CD, the data are normalized so that each coefficient has a mean and unit variance of zero. RBMs have also been used as density models to represent the distributions of acoustic features, and can be extracted as a low-dimensional feature in hidden layers.

2.2.3 | Acoustic speaker modeling

By using feature vectors extracted from a given speaker's training utterances, a speaker model is trained and stored in the system database. The feature extraction and speaker modeling steps jointly represent the training or enrollment phase of an SI in which speakers are registered in the SI system. In the TD mode, the model is utterance specific, and includes the temporal dependencies between the feature vectors. In the TI mode, the feature distribution is often modelled, rather than including the temporal dependencies. The goal of this stage is to develop unique templates or models for each enrolled speaker.

There are various acoustic modeling methods for SI systems, mainly the Gaussian mixture model (GMM) and iVector [22]. In this study, we performed speaker modeling based on artificial neural networks (ANNs) using MLP networks. The details of the proposed structure are given in Section 3.

2.2.4 | Classification

In this stage, an unknown voice is used as an input to the system to return the most relevant speaker name/identifier. This classification is collectively referred to as the testing phase in which the SI system is evaluated on the basis of its classification accuracy. Pattern matching is entirely dependent on the nature of the acoustic speaker models. In the case of stochastic generative models, matchings are quantified in the form of log-likelihood scores, whereas for parametric ones, they may be simple distance metrics. For discriminative models, scores may be based on the distance from the decision boundary of

two classes (for example, in support vector machines) or the difference between the actual and predicted class (for example, in ANNs). As we have used ANNs for acoustic modeling in this work, the classification was also performed using this technique.

3 | PROPOSED METHOD

In this paper, we propose the use of ANNs to learn DBN-based features that are used in developing a TD and TI dysarthric speaker identifier. In this section, the structure of the algorithm and parameters of the proposed method are presented.

The acoustic feature set that is extracted from each frame of the speech signals is a 39-dimensional vector consisting of 12 MFCC, 1 log-energy, 13 delta features, and 13 delta-delta features. The acoustic features are stacked and pass through a DBN to yield a new representation of features. For a consistent comparison of our proposed features with the standard MFCC, we extracted 39 features from the DBN.

In the proposed DBN-based feature extractor, as shown in Figure 3, the DBN structure is a three-layer network consisting of 195 input neurons, one hidden layer with a size of 500 neurons, and an output layer with 39 neurons. To feed a 39-dimension MFCC vector into the network, the four vectors next to the input vector are concatenated and given as an input to the network. This is done in order to model the dynamic effect of consequence frames.

We have proposed two scenarios for acoustic modeling using MLP neural networks:

1. Single-network classifier: In this structure, an MLP network with 39 input neurons, a hidden layer, and an output layer is used, where the number of output neurons is equal to the number of speakers.
2. Multi-network classifier: In this scenario, dysarthric speakers are categorized into three groups based on their dysarthria severity or speech intelligibility, and a 3-layer MLP is trained for each category. The first category of speakers consists of dysarthric speakers with an intelligibility rate of 2%–34%; we call this group *High Severity*. The second category consists of speakers with a speech intelligibility of 35%–62%, and is called *Mid Severity*. Finally, the third category is called *Low Severity*, and comprises individuals with a speech intelligibility rate of 63%–95%.

In each classifier in single-network, multi-network, and baseline system ANN-based acoustic modeling, we used 3-layer MLPs with 117 input neurons (covering three consequent DBN-based features) and 1,000 hidden neurons. The number of output neurons was equal to the number of speakers. In addition, the learning rate was set to 0.001. The sigmoid function was used as the activation function of the hidden layer, and the soft-max function was used as the activation function of the output layer. We used 3-fold cross-validation in the training phase.

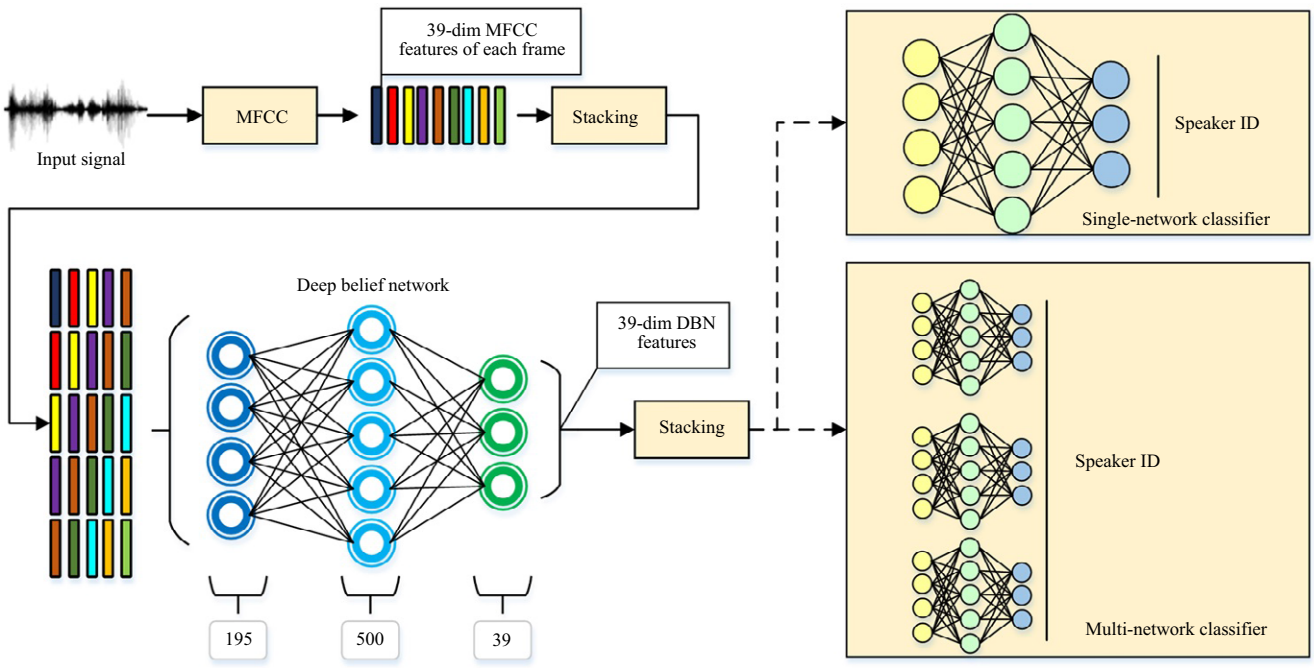

FIGURE 3 Proposed DBN-MLP based speaker identification structure.

4 | EXPERIMENTS AND RESULTS

4.1 | Dataset

In our experiments and evaluations, we used speech data provided by the UA-Speech database of dysarthric speakers from the University of Illinois [8]. This dataset was collected from 19 male and female speakers with different dysarthric diagnoses and different severity levels of dysarthria, varying from very low speech intelligibility (2%) to high intelligibility (95%) (Table 1). This database also contains the speech utterances of 11 normal speakers (that is, speakers without disabilities) that were collected under the same circumstances and with the same vocabulary used as control speakers. Therefore, these data allow us to compare the performance of the same speaker identification system for both dysarthric and normal speakers. This database consists of 255 isolated words per speaker, including 19 computer commands (for example, "Cut," "Past"), 100 common words (for example, "Like," "Go"), 10 digits (0–9), 26 radio alphabet letters (for example, "Alpha," "Bravo"), and 100 uncommon words (for example, "naturalization," "moonshine"). For each word, three repetitions were recorded, and for each repetition, seven channels of speech were saved as seven separate wav files that were presented as Microsoft PCM at a sampling rate of 16 kHz.

The categorization on this database enables its use in speaker recognition tasks. In this study, we used voices of 16 dysarthric speakers (12 male and 4 female) given in Table 1 for speaker identification. In the training phase, approximately 5 minutes of speech data was used for each speaker, and 100 short utterances (each utterance is approximately 4 seconds) were used for the testing phase in both TD and TI scenarios. For each scenario, experimental results for the two test sets are reported.

TABLE 1 Summary of speakers' information obtained from UA speech database for use in this paper

No.	Speaker ID	Sex	Age	Speech intelligibility	Dysarthria diagnosis
1	F02	Female	30	Low (29%)	Spastic
2	F03	Female	51	Very low (6%)	Spastic
3	F04	Female	18	Mid (62%)	Mixed
4	F05	Female	22	High (95%)	Spastic
5	M01	Male	>18	Very low (15%)	Spastic
6	M04	Male	>18	Very low (2%)	Spastic
7	M05	Male	21	Mid (58%)	Spastic
8	M06	Male	18	Low (39%)	Spastic
9	M07	Male	58	Low (28%)	Spastic
10	M08	Male	28	High (93%)	Spastic
11	M09	Male	18	High (86%)	Spastic
12	M10	Male	21	High (93%)	Mixed
13	M11	Male	48	Mid (62%)	Ataxic
14	M12	Male	19	Very low (7.4%)	Mixed
15	M14	Male	40	High (90.4%)	Spastic
16	M16	Male	–	Low (43%)	Spastic

4.2 | Dysarthric speaker identification: Baseline

As previously mentioned, dysarthria is a degenerative disorder of the central nervous system, and reduces control over articulation and pitch [14]. This affects the uniqueness of sounds produced by dysarthric people, which impairs the performance of conventional speaker recognition systems. The first set of experiments was designed to demonstrate the applicability of the baseline speaker recognition system for dysarthric speakers. The baseline system of this work uses 39 MFCC features as the feature vector of each frame, and a 3-layer MLP neural network as the classifier.

Using baseline MFCC features, two MLPs, one for normal and another for dysarthric speakers, were used for classification. In both cases, eight speakers were used. For a fair comparison, all parameters in both evaluations were the same. Table 2 shows the accuracy of this speaker identification system. Nevertheless, there is no linear correlation between the speaker identification accuracy and the intelligibility level for dysarthric speakers. However, it is evident that for dysarthric speakers, the system performance is unreliable, especially for speakers with high dysarthria severity.

4.3 | DBN-based speaker identification

In this section, we present our experimental results obtained for the proposed dysarthric speaker identification

TABLE 2 Speaker identification performance for normal and dysarthric speakers using MFCC features (Baseline)

Dysarthric speakers			Normal speakers		
Speaker ID	Intellig. (%)	Accuracy (%)	Speaker ID	Intellig. (%)	Accuracy (%)
F02	29	98	CF02	100	96
F03	6	92	CF03	100	100
F04	62	100	CF04	100	98
F05	95	98	CF05	100	99
M01	15	55	CM01	100	100
M04	2	85	CM04	100	96
M05	58	100	CM05	100	98
M06	39	100	CM06	100	99
Average		91.00	Average		98.25

TABLE 3 Performance of text-dependent single-network system (TD-SN)

	Speaker properties		TD speaker identification rate	
No.	Intellig. (%)	ID	MFCC features (%)	DBN features (%)
1	29	F02	100	99
2	6	F03	100	100
3	62	F04	98	100
4	95	F05	100	95
5	15	M01	90	98
6	2	M04	53	55
7	58	M05	86	93
8	39	M06	100	100
9	28	M07	99	99
10	93	M08	87	94
11	86	M09	95	98
12	93	M10	73	80
13	62	M11	62	94
14	7	M12	100	100
15	90	M14	73	90
16	43	M16	98	94
Average			88.37	93

TABLE 4 Performance of text-independent single-network system (TI-SN)

	Speaker properties		TI speaker identification rate	
No.	Intellig. (%)	ID	MFCC features (%)	DBN features (%)
1	29	F02	97	91
2	6	F03	50	72
3	62	F04	95	98
4	95	F05	98	93
5	15	M01	85	94
6	2	M04	36	41
7	58	M05	82	90
8	39	M06	99	99
9	28	M07	90	96
10	93	M08	60	70
11	86	M09	56	71
12	93	M10	45	51
13	62	M11	53	73
14	7	M12	100	99
15	90	M14	49	61
16	43	M16	83	83
Average			73.6	80.1

systems presented in Section 3. In the implementation of DBN, we used the MATLAB toolbox for DBN, which is called DeeBNet[1)] [23].

For comparison, the experiments of the dysarthric SI system are performed using both MFCC and DBN features. The evaluations were performed for the four following scenarios:

- Text-Dependent Single-Network (TD-SN)
- Text-Independent Single-Network (TI-SN)
- Text-Dependent Multi-Network (TD-MN)
- Text-Independent Multi-Network (TI-MN)

In both TD and TI systems, 100 short utterances or words were used for the test phase. Accordingly, in TD mode, the utterances in the test and train phases are similar, while in TI mode, the utterances are completely different. The recognition accuracy results of the TD-SN system are shown in Table 3, and for the TI-SN system, the results are presented in Table 4. It is obvious that in both TD and TI cases, the accuracy rate of DBN-based systems is higher than MFCC for most speakers. Further, as expected, the performance of the TD system is higher than that of the TI one. The improvement, which is caused by using DBN features id, is 4.6% and 6.5% for the TD and TI systems, respectively. These results confirm the robustness of DBN features compared with MFCC in the presence of the signal variability of dysarthric speakers.

Tables 5 and 6 list the comparative accuracy percentages of the MFCC and DBN features in TD-MN and TI-MN testing sets, respectively. It can be seen that similar to the single-network cases, the DBN features resulted in higher performance than the MFCC. The best average accuracy rate in these experiments is 97.3%, which is achieved by the DBN-based system in the TD case. A comparison of the performance of multi-networks with single-network systems shows the effectiveness of the multi-network structure. However, it should be noted that for the real-word implementation of multi-network structures, the dysarthria severity of the speaker should be determined or calculated.

To obtain a general comparison of the proposed structures, the average speaker recognition accuracy rates of MFCC and DBN features are also given in Figure 4. In this figure, that the following observations can be made:

- DBN-based features are superior to MFCC in all experiments. This confirms the robustness of the DBN representation, especially for variant acoustic conditions such as in dysarthric speech.

[1)] http://ceit.aut.ac.ir/~keyvanrad/

TABLE 5 Performance of text-dependent multi-network system (TD-MN)

Category	Speaker properties		TD speaker identification rate	
	Intellig. (%)	ID	MFCC features (%)	DBN features (%)
High severity	2–34	F02	100	100
		F03	92	90
		M01	93	96
		M04	79	86
		M07	89	97
		M12	100	100
Mid severity	35–62	F04	100	100
		M05	98	100
		M06	100	100
		M11	92	96
		M16	97	98
Low severity	63–95	F05	99	100
		M08	100	100
		M09	100	100
		M10	97	99
		M14	92	95
Average			95.5	97.3

TABLE 6 Performance of text-independent multi-network system (TI-MN)

Category	Speaker properties		TI speaker identification rate	
	Intellig. (%)	ID	MFCC features (%)	DBN features (%)
High severity	2–34	F02	100	99
		F03	52	50
		M01	82	95
		M04	69	83
		M07	60	63
		M12	100	100
Mid severity	35–62	F04	100	90
		M05	99	99
		M06	100	100
		M11	78	82
		M16	89	91
Low severity	63–95	F05	98	99
		M08	99	99
		M09	99	99
		M10	75	92
		M14	89	92
Average			86.8	90.12

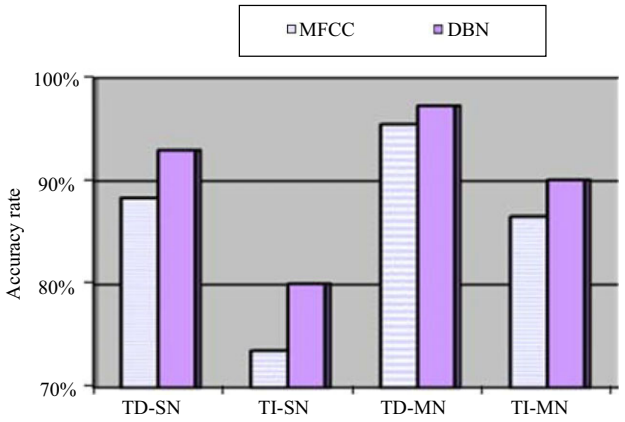

FIGURE 4 Performance of proposed systems compared with the baseline

- The proposed multi-network (MN) structure results in higher accuracy compared with the SN structure. This is because in the MN case, each network learns a simpler classification task than in the SN case.
- Similar to the case for normal speakers (and as expected), the TD recognition rate is higher than the TI rate for dysarthric speakers.

5 | SUMMARY AND CONCLUSIONS

In this work, we proposed a dysarthric speaker identification system using ANNs by applying DBN for signal representation and MLP for classification. For classification, the two different structures, SN and MN, were proposed for the MLP neural network.

As the first study in speaker identification for dysarthric speakers using the UA database, our evaluations used 100 utterances for each speaker from the UA-speech database. The evaluations were done for TD and TI tasks, and the proposed results of the proposed DBN-based feature-recognition method were compared with the MFCC features. Our experimental results demonstrated that the proposed DBN-based features have good robustness compared with the MFCC feature. The best setup of our proposed systems (DBN-based features, TD, and MN MLP) achieved a speaker recognition rate of 97.3%.

ORCID

Aref Farhadipour http://orcid.org/0000-0002-3447-4330

REFERENCES

1. F. Rudzicz, *Production knowledge in the recognition of dysarthric speech*, Ph.D. Thesis, Dept. Comput. Sci, Toronto University, Canada, 2011.
2. V. Poblete et al., *A perceptually-motivated low-complexity instantaneous linear channel normalization technique applied to speaker verification*, Comput. Speech Lang. **31** (2015), no. 1, 1–27.
3. M. J. Kim, Y. Kim, and H. Kim, *Automatic intelligibility assessment of dysarthric speech using phonologically-structured sparse linear model*, IEEE/ACM Trans. Audio, Speech, Lang. Process. **23** (2015), no. 4, 694–704.
4. B. Schuller et al., *A survey on perceived speaker traits: Personality, likability, pathology, and the first challenge*, Comput. Speech Lang. **29** (2015), no. 1, 32.
5. K. L. Kadi et al., *Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge*, Biocybernetics Biomed. Eng. **36** (2016), no. 1, 233–247.
6. X. Menendez-Pidal et al., The nemours database of dysarthric speech, *Proc. Int. Conf. Spoken Lang.*, Philadelphia, PA, USA, Oct. 3–6, 1996, pp. 1962–1965.
7. F. Rudzicz, A. K. Namasivayam, and T. Wolff, *The TORGO database of acoustic and articulatory speech from speakers with dysarthria*, Lang. Resour. Eval. **46** (2012), no. 4, 523–541.
8. H. Kim et al., *Dysarthric speech database for universal access research*, Interspeech **2008** (2008), 1741–1744.
9. S. R. Shahamiri, B. Salim, and S. Salwah, *A multi-views multi-learners approach towards dysarthric speech recognition using multi-nets artificial neural networks*, IEEE Trans. Neural Syst. Rehabil. Eng. **22** (2014), no. 5, 1053–1063.
10. S.-O. Caballero-Morales and F. Trujillo-Romero, *Evolutionary approach for integration of multiple pronunciation patterns for enhancement of dysarthric speech recognition*, Expert Syst. Applicat. **41** (2014), no. 3, 841–852.
11. S. R. Shahamiri and S. S. B. Salim, *Artificial neural networks as speech recognisers for dysarthric speech: Identifying the best-performing set of MFCC parameters and studying a speaker-independent approach*, Adv. Eng. Inform. **28** (2014), no. 1, 102–110.
12. G. Hinton et al., *Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups*, IEEE Signal Process. Mag. **29** (2012), no. 6, 82–97.
13. Z.-H. Ling et al., *Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends*, IEEE Signal Process. Mag. **32** (2015), no. 3, 35–52.
14. F. Rudzicz, *Articulatory knowledge in the recognition of dysarthric speech*, IEEE Trans. Audio Speech Lang. Process. **19** (2011), no. 4, 947–960.
15. R. Palmer and P. Enderby, *Methods of speech therapy treatment for stable dysarthria: A review*, Int. J. Speech-Lang. Pathol. **9** (2007), no. 2, 140–153.
16. T. Kinnunen and L. Haizhou, *An overview of text-indepedent speaker recognition from features to supervectores*, Speech Commun. **52** (2010), no. 1, 12–40.
17. X.-L. Zhang and J. Wu, *Deep belief networks based voice activity detection*, IEEE Trans. Audio Speech Lang. Process. **21** (2013), no. 4, 697–710.
18. J. Sohn and W. Sung, A voice activity detector employing soft decision based noise spectrum adaptation, *IEEE Int. Conf. Acoustics, Speech, Signal Process.*, Seattle, WA, USA, May 15, 1998, pp. 365–368.
19. S. B. Davis and P. Mermelstein, *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*, IEEE Trans. Acoust. Speech Signal Process. **28** (1980), no. 4, 357–366.
20. H. Hermansky, *Perceptual linear predictive (PLP) analysis of speech*, J. Acoust. Soc. Am. **87** (1990), no. 4, 1738–1752.
21. G. E. Hinton, *Training products of experts by minimizing contrastive divergence*, Neural Comput. **14** (2002), no. 8, 1771–1800.
22. N. Dehak et al., *Front-end factor analysis for speaker verification*, IEEE Trans. Audio Speech Lang. Process. **19** (2011), no. 4, 788–798.
23. M. A. Keyvanrad and M. M. Homayounpour, *A brief survey on deep belief networks and introducing a new object oriented MATLAB toolbox (DeeBNet)*, arXiv preprint arXiv:1408.3264, 2014.

AUTHOR BIOGRAPHIES

Aref Farhadipour received his BS degree in electronic engineering from LIA University, Lahijan, Iran in 2013, and his MS degree in sound engineering from the IRI Broadcast University, Tehran, Iran, in 2015. Since 2017, he has been a staff member at an Iranian broadcast organization. His research interests include speech processing, deep learning, image processing, and pattern recognition.

Hadi Veisi received his PhD in artificial intelligence from the Sharif University of Technology, Tehran, Iran in 2011. He joined the Faculty of New Sciences and Technologies at the University of Tehran in 2012, and established the Data and Signal Processing laboratory. His primary research interests are artificial neural networks and deep learning, natural language processing, and speech processing.

Mohammad Asgari received his BS degree in telecommunication and electronic engineering from the University of Science and Technology, Tehran, Iran in 1993, and his MS and PhD degrees in telecommunication and signal processing from the Iran University of Science and Technology, Tehran, Iran in 1997 and 2002, respectively. Since 1994, he has been a staff member in the technical department of an Iranian broadcast organization. His research interests include speech processing, microphone arrays, and audio watermarking.

Mohammad Ali Keyvanrad received his BS degree in software engineering from the Amirkabir University of Technology, Tehran, Iran, in 2007, and his MSc and PhD degrees in artificial intelligence from the Amirkabir University of Technology, Tehran, Iran, in 2010 and 2016, respectively. His research interests include pattern recognition, machine learning and signal processing, especially deep learning, feature learning, and audio indexing.