Advanced Engineering Informatics 28 (2014) 102–110

Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei

Artificial neural networks as speech recognisers for dysarthric speech: Identifying the best-performing set of MFCC parameters and studying a speaker-independent approach

Seyed Reza Shahamiri *, Siti Salwah Binti Salim [1]

Department of Software Engineering, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia

ARTICLE INFO

Article history:
Received 10 April 2013
Received in revised form 6 January 2014
Accepted 11 January 2014

Keywords:
Dysarthria
Automatic speech recognition
Artificial neural network
Mel-frequency cepstral coefficients

ABSTRACT

Dysarthria is a neurological impairment of controlling the motor speech articulators that compromises the speech signal. Automatic Speech Recognition (ASR) can be very helpful for speakers with dysarthria because the disabled persons are often physically incapacitated. Mel-Frequency Cepstral Coefficients (MFCCs) have been proven to be an appropriate representation of dysarthric speech, but the question of which MFCC-based feature set represents dysarthric acoustic features most effectively has not been answered. Moreover, most of the current dysarthric speech recognisers are either speaker-dependent (SD) or speaker-adaptive (SA), and they perform poorly in terms of generalisability as a speaker-independent (SI) model. First, by comparing the results of 28 dysarthric SD speech recognisers, this study identifies the best-performing set of MFCC parameters, which can represent dysarthric acoustic features to be used in Artificial Neural Network (ANN)-based ASR. Next, this paper studies the application of ANNs as a fixed-length isolated-word SI ASR for individuals who suffer from dysarthria. The results show that the speech recognisers trained by the conventional 12 coefficients MFCC features without the use of delta and acceleration features provided the best accuracy, and the proposed SI ASR recognised the speech of the unforeseen dysarthric evaluation subjects with word recognition rate of 68.38%.

1. Introduction

Dysarthria is a neurological impairment that damages the control of the motor speech articulators, which the malfunction is caused by the lack of control over the speech-related muscles, the lack of coordination among them, or their paralysis. It is often associated with irregular phonation and amplitude [1,2]. As a result of the impairment, the speech signal is compromised and its intelligibility is reduced [3,4]. According to [5], low intelligibility is one of the most detrimental social characteristics of dysarthria that affects different aspects of the lives of people with such disability.

Automatic Speech Recognition (*ASR*) systems identify the uttered word(s) represented as an acoustic signal and rely on a given lexicon to recognise the spoken word(s). They have several applications in health care, the military, telephony, and other domains [6]. They can be very helpful for speakers with dysarthria, because the disabled persons are often physically incapacitated and unable to use keyboards [7,8].

Most state-of-the-art commercial ASR systems are designed for speakers without speech disabilities, (i.e. non-speech disordered) and exclude those with speech disabilities [9]. These ASR systems record lower performance for individuals who suffer from dysarthria (specifically severe dysarthria [10,11]) than for people without speech disabilities as dysarthric speech is different from normal speech [12–14]. Therefore, there has recently been a trend towards creating specialised ASR systems for individuals with dysarthria instead of using ASR systems designed primarily for speakers without speech disabilities [3,10,15,16]. Thus, it is necessary to propose an ASR model specifically built for users with dysarthria that delivers adequate accuracy; specialised systems have generally achieved comparatively better performance for people with speech disorders [2,4,10,16].

According to [9], it is easier for people with dysarthria to utter isolated words rather than a continuous sequence of words. Similarly, it is more effective when the size of the ASR vocabulary is small and includes only simple words with one or two syllables in order to boost recognition rates with reduction or minimisation of dysarthric ASR errors [9]. Therefore, isolated-word and small-vocabulary ASR models are in greater demand for dysarthric speech recognition [10,15].

* Corresponding author. Address: BS16, Block B, Level 1, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia. Tel.: +60 3 79676300; fax: +60 3 79579249.

E-mail addresses: admin@rezanet.com (S.R. Shahamiri), salwa@um.edu.my (S.S. Binti Salim).

[1] Address: Faculty of Computer Science and Information Technology, University of Malaya, Lembah Pantai, 50603 Kuala Lumpur, Malaysia.

1474-0346/$ - see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.aei.2014.01.001

Although ASR technologies for dysarthria have been considered, previous studies show that ASR systems for users with dysarthria have not yet attained an adequate performance level in terms of generalisability because of the complex issues related to dysarthric speech [13]. For example, increased variability due to physical fatigue and frustration of individuals with dysarthria, as well as variations in the severity levels of the disease, make it difficult to produce an ASR model to be used by most individuals with dysarthria.

A speaker-dependent (SD) ASR system is capable of recognising the speech of users whose acoustic data have been captured while training the system [17]. If an unknown speaker uses the system, the accuracy of the system is reduced. In the context of dysarthric ASR system, as the performance of an ASR system is reduced without proper training data, a major problem with SD paradigm is that people with dysarthria may be rapidly fatigued by the effort needed to provide the vocabulary in order to train the ASR system [18,19].

Dysarthric speaker-adaptive (SA) ASR models are usually trained as normal speaker-independent (SI) ASR models, but they adapt to new users' data while the systems are being used by the disabled people. Particularly, the systems learn the speech of new users every time they utter a new word. These systems may provide low recognition rates during early stage of usage, but the performance will gradually improve over longer time of usage [20]. Hence, they do not recognise dysarthric speech properly out-of-the-box. Furthermore, from the perspective of performance, none of these models are capable of identifying speech uttered by unforeseen users accurately, and therefore, they are not suitable for applications such as in the public telephone network [17].

In contrast, SI speech recognisers are trained with databases containing utterances of several speakers. These systems are capable of recognising the speech of a variety of users more accurately than speaker dependent or adaptive ASR systems, including the speech of users whose acoustic data have not been provided during the training process. Consequently, an SI ASR model for users with dysarthria is required in order to recognise more accurately the speech of a wide range of users with speech disabilities; such ASR system can then be accessed by speech-disabled people using public services [11]. As an illustration, usage of banking phone services requires a user to input menu commands by pressing physical numeric buttons located on a phone. Normal people can easily press the keypad buttons, but dysarthric people may be unable to do that, because they are usually physically handicapped. As such, an ASR system is extremely useful in facilitating the disabled people to utter the numbers associated with the menu commands. SD or SA ASR models are incapable of performing this function with sufficient accuracy for new users, but SI ASR systems may be capable of providing the required generalisability and performance so that speech-disabled people can communicate their instructions by using public services. Therefore, building SI ASR systems designed for users with dysarthria is an important topic that should be investigated further.

In order for an ASR system to be operable, acoustic features of utterances must be presented to the system using a process called *Feature Extraction*. The usage of *Mel-Frequency Cepstral Coefficients* (*MFCCs*) is the most common feature-extraction method in ASR applications, which represent speech signals in cepstral domain [20]. It is a representation defined as the real cepstrum of a windowed short-time signal derived from the Fast Fourier Transform of that signal, in which the frequency bands are spaced on the mel scale equally (inspired by the human auditory perception system). MFCCs have been widely used for several speech-disorder signal-processing tasks such as speech disability classification [21,22] and dysarthric speech recognition. The MFCCs are usually presented as mel cepstrum with 12 coefficients, their first and second derivatives.

MFCCs have been proven to be an appropriate representation of dysarthric speech [3,23,24]. Although it is advisable to use all MFCC-based feature sets (i.e. MFCCs (12 coefficients), Delta-MFCCs, and Delta/Delta MFCCs and log energies) together for training an ASR system for normal speakers, it remains unexplored if all the MFCC features and/or its combination with the first and second derivative should be used as inputs to dysarthric ASR systems based on *Artificial Neural Networks* (*ANNs*). This question should be looked into because normal speech is different from dysarthric speech, and the same goes for the respective acoustic features. Selecting the best representation set of dysarthric acoustic features is a crucial issue because it may directly influence the recognition accuracy of dysarthric ASR systems.

Hence, this paper attempts to resolve the above issues relating to dysarthric ASR systems. The objectives of this study are:

1. To identify the most effective MFCC-based feature set for representing dysarthric acoustic signals in order to provide an ANN-based dysarthric ASR model. The MFCC parameters considered here are mel cepstrum with 12 coefficients, their first and second derivatives, and all the acoustic features.
2. To study the application of ANNs in a fixed-length, isolated-word SI dysarthric ASR system. The vocabulary size is 11 including 10 digits and silence.

The first objective was achieved by providing 28 SD ANN-based ASR systems over seven dysarthric subjects, the results of which were compared. For each speaker, four speech recognisers were provided, and each of them was trained by one set of the MFCC parameters (12 MFCCs, Delta-MFCCs, Delta-Delta MFCCs, and all sets together). The energy information was not considered here, because it is often difficult for dysarthric individuals to maintain a steady volume; hence, according to Green et al. [18], the energy information may not be useful. The second objective was accomplished by proposing an automatic dysarthric digit recogniser, which provides spoken-numerical-command capability; this model is very useful for users with speech disabilities who are physically incapacitated. The accuracy and recognition rate of the proposed SI ASR system were measured by using evaluation data collected from people with severe, moderate, and mild dysarthria respectively. It is pertinent to note that no acoustic sample from the evaluation subjects was considered during the SI ASR training in order to highlight the generalisability and speaker independency of the proposed ASR model.

2. Previous works

This section surveys the studies of state-of-the-art ASR technologies for users with dysarthria. Most of the experimental dysarthric speech recognition systems resorted to SD or adaptive approaches because of the above-mentioned dysarthric speech issues. As an illustration, Hasegawa-Johnson et al. [16] provided two isolated-word SD ASR systems (10-digit vocabulary) based on the data collected from three subjects with dysarthria: one female and two males with one control subject. The speech samples were recorded using an array of seven microphones and four cameras mounted on top of a computer monitor. The first system was a phone-based *Hidden Markov Model* (*HMM*), and the second was a fixed-length isolated-word ASR system based on *Support Vector Machines* (*SVMs*). The former was successful for two subjects, but it failed for one of the subjects with the tendency to delete consonants in a word. Similarly, the SVM-based ASR failed to perform for one of the subjects with dysarthria, because he suffered from

stuttering, but it was successful for the other two subjects. The authors concluded that HMM-based dysarthric ASR models may provide robustness against large-scale word-length fluctuations, and SVM-based models can handle the deletion or reduction of consonants.

Selouani et al. [3] proposed another SD ASR based on HMMs for English and French speakers with dysarthria for continuous speech. The ASR system was trained using speech materials collected from four dysarthric speakers in the Nemours database and one control speaker; the authors did not mention the severity of the subjects' disabilities. The training speech samples were presented by mel cepstrum with 12 coefficients, their first and second derivatives, and their log energies. The training set is composed of 50 sentences (300 words), and the test is composed of 24 sentences (144 words). The average recognition rate of this SD system was 70% for the four dysarthric subjects.

STARDUST, a HMM-based ASR system for users with severe dysarthria, was introduced in [10,18]. In this system, a new HMM was trained every time the user uttered a new word. The training and evaluation data were obtained from five individuals with dysarthria and were presented as mel cepstrum with 12 coefficients inclusive of their first derivatives. The speech samples were collected by Andrea DA.400 microphone array or Acoustic Magic Voice Tracker array at distances of 0.5–3 m from the participants. The system was an isolated-word ASR and included a 10-digit vocabulary. In another example of isolated-word HMM-based ASR for users with dysarthria [25], a small and medium vocabulary size (SD) ASR system for spastic dysarthria was studied.

ANN-based ASR system have been successfully employed for normal speech as reported in the literature (such as [26–28]). ANNs are mathematical models inspired by natural neural systems that learn the function by capturing information from given input and output samples. Jayaram and Abdelhamied studied the application of ANNs in a SD dysarthric speech recognition system but with limited success [29]. The authors applied ANNs in a 10-word ASR system to recognise the speech of one speaker with severe dysarthria. They provided two recognisers: the first was trained using MFCC parameters and the second using the formant frequencies; it was found that the first system performed better than the second and outperformed five human listeners. The results of this study are not concrete, because the ASR system was trained and evaluated using only one subject with dysarthria.

Several studies on ANN/HMM hybrid ASR showed that the hybrid model is a suitable platform for normal ASR [26]. However, the applications of hybrid ASR for users with dysarthria have not been widely studied, because proper ANN training data to perform dysarthric phoneme recognition are not easily achievable. Neural networks within the ASR hybrid approach (normal speakers) are usually applied to provide the language model since the hybrid approach is phone-based; nonetheless, for dysarthric speech this is a challenging task, because identifying the phones and labelling them, i.e. segmenting the speech utterances for dysarthria, is a difficult, error-prone, and time-consuming process due to low speech intelligibility of the disabled persons. Moreover, since the unintelligibility of dysarthric speech is because of the combination of many articulatory behaviours that can lead to phonemic insertion errors in or around words [30,31], dysarthric ASR approaches that consider word-based units may be more successful than those which depend on phone based units.

Despite the above SD and SA systems, there had been a few unsuccessful attempts to provide SI ASR systems for users with dysarthria. They were unsuccessful because the error rates were too high for these speech recognition systems to be of any practical application. For example, Sanders and his colleagues [13] studied how a normal, SI HMM-based ASR system behaved when it was used by people with dysarthria. The ASR, trained with non-speech-disordered speech data, was evaluated with dysarthric data acquired from two male speakers with mild dysarthria. The results for the two evaluation subjects showed WER of 15.4% and 41% respectively. However, the same ASR system had better performance when it was trained and tested with dysarthric data, (i.e. as a SD ASR). The SD ASR system for the same two dysarthric subjects had WER of 2.6% and zero respectively. Similar results were described by Talbot for the ENABL project [14]. The author verified a commercial ASR system with data collected from 10 individuals with dysarthria (five males and five females) and reported that the error rate was as high as 71%.

Sharma and Hasegawa-Johnson considered the database used in this study to provide two isolated-word HMM-based speech recognisers for users with dysarthria (one SD and another SA) [19]. For the adaptive model, they provided an isolated-word, SI ASR system for speakers without disabilities first; the system was based on the TIMIT database, in which *Perceptual Linear Prediction* (*PLP*) coefficients were extracted as acoustic features. Subsequently, the authors utilised the speech of seven speakers with dysarthria from the UA-Speech database to verify the normal ASR as a SA dysarthric ASR system. The maximum average recognition rate for the SA systems was 36.8% and 30.84% for the SD systems. However, to the best of our knowledge, they did not provide any SI model for speakers with dysarthria. The method proposed here achieved better results even as an SI ASR system.

Therefore, the literature review shows that there is no SI ASR model designed specifically for individuals with dysarthria. This study explains how ANNs can be trained by word-based acoustic features to be used as a fixed-length automatic digit recogniser, which has the capability of recognising speech of unknown dysarthric individuals. Such speaker-independent ASR system will benefit a wider range of people with dysarthria. Moreover, each of the previous studies considered a different set of acoustic features. Therefore, it is important to identify which MFCC-based feature set represents dysarthric acoustic features most effectively.

3. Methods

3.1. Materials and participants

A few dysarthric databases were available at the time of writing this report. However, most of them were not suitable for this research since the context of this study was to provide a fixed-length, isolated-word digit speech recogniser. Thus, an isolated-word dysarthric speech corpus including enough utterances of the vocabulary, (i.e. digits 0–9) was necessary. In addition, training a speaker-independent ASR system requires a large number of participants in order to increase the generalisability of the classifier. Hence, we used speech materials provided by the UA-Speech Database for dysarthria, which was produced by the University of Illinois [32]. The database contains isolated-words, acoustic samples of digits, radio alphabet letters, computer commands, and common words acquired from 19 male and female subjects with dysarthria of different severity levels, varying from extremely low speech intelligibility (2%) to high intelligibility (95%). The speech data were recorded at a sampling rate of 48 kHz using an eight microphones (6 mm in diameter) array with 1.5 in. of spacing between adjacent microphones. The array was mounted at the top of the laptop computer screen next to a video camera used to capture the visual features of speech. Each utterance contains only a single word so that no word detection module was necessary.

The vocabulary size of the database is 455 but we only utilised the 10-digit utterances of 16 of the subjects with dysarthria to provide the required speech materials for ASR modelling and evaluation. For each utterance, there are three examples of each digit

per subject. Table 1 provides more information about the subjects with dysarthria used in this research. As can be seen, all dysarthric severity classes among male and female participants are covered in order to highlight the capability of the proposed ASR system as an SI paradigm. Moreover, acoustic samples of 11 speakers without disabilities for the same vocabulary (provided by the database) are also considered as control speakers; this will provide a benchmark for measuring the performance of the proposed system when it evaluates data given by normal speech.

Speech therapists often use clinical assessments of intelligibility in dysarthric speakers for rehabilitation [33]. Speech intelligibility is the measure of the degree of dysarthric speech understandability, and it correlates well with accuracy of SD, SI, or SA dysarthric ASR systems [8]. UA-Speech database presents dysarthric speech severity by the speech intelligibility of each speaker. Here we classified the speech severity as high, moderate, or mild based on the participants' intelligibility provided by the database. If the speech is identified as "High Dysarthric Severity", its intelligibility is low (less than 33%). On the other hand, "Mild Dysarthric Severity" means the intelligibility is high (between 66% and 99%). The rest of the intelligibility values, ranging between 33% and 66%, are defined as "Moderate Dysarthric Severity". In this study, we measured the performance of the proposed SI ASR model for each of the above severity levels separately.

3.2. The ANN-based ASR model for users with dysarthria

The first step in building an ANN-based ASR system is to train it with speech utterances. In order to train the ANN, speech samples must be represented as acoustic features; these features are usually provided by the feature extraction procedure. After several experiments with different frame and sliding window sizes, the best performance was provided by selecting 22 frames of MFCC features for each digit utterance (each utterance is an isolated digit), in which the frame size was 162.5 ms with a sliding hamming window of 81 ms (i.e. 50% frame overlapping). The number of frames was selected to match the maximum length of the utterances (1782 ms) provided by the database. The silence at the beginning and the end of each utterance was removed to ensure that each utterance commenced with useful acoustic data instead of silence. Nevertheless, for utterances smaller than 22 frames, the missing frames were replaced by silence frames at the end of the utterance, in order to solve the length variability issue of dysarthric speech. The features extracted from each frame were concatenated to the previous frames to create the ANN input vector. As an illustration, for the experiments conducted with mel cepstrum with 12 coefficients, each utterance was represented by a vector of 264 features (12 features per frame $\times$ 22 frames); each feature was assigned to one of the input neurons, (i.e. each 12-input neuron represented one of the frames). When the features were ready, the ANN was trained with the extracted word-feature vectors.

We considered MLP neural networks with three layers. The input layer should have one neuron for each speech feature. The output layer must have one neuron for each item in the vocabulary. The feed-forward and back-propagation training procedure was selected as the training algorithm. The rest of the ANN parameters are explained in the next section.

Once the utterance of a disabled person is given to the system, the same feature extraction procedure must be applied to the utterance before it is fed to the ANN. Next, the extracted features are provided to the trained ANN as an input vector. Each of the output neurons produces a result for the given input vector, and the one with the maximum value is identified as the recognised digit. The entire process of training and using the ANN-based ASR system for users with dysarthria is shown in Fig. 1.

Table 1
The training and evaluation subjects with dysarthria [32].

Participant	Sex	Age	Severity of dysarthria
M01	Male	>18	High
M04	Male	>18	High
M05	Male	21	Moderate
M06	Male	18	Moderate
M07	Male	58	High
M08	Male	28	Mild
M09	Male	18	Mild
M10	Male	21	Mild
M11	Male	48	Moderate
M12	Male	19	High
M14	Male	44	Mild
M16	Male	40	High
F02	Female	30	High
F03	Female	51	High
F04	Female	18	Moderate
F05	Female	22	Mild
Control speakers			CF02[a], CF03, CF05, CM01[a], CM04, CM05, CM06, CM08, CM09, CM12, CM13

[a] CF is a female subject and CM is a male subject.

3.3. Evaluation criteria

Accuracy and word recognition rate are considered as the evaluation criteria in order to assess the quality of the ANN-based speech recognisers produced in this study. These two parameters are defined as follows [28]:

1. *Word Recognition Rate (WRR):* The proportion of correct identification of the words, (i.e. digits) by the ASR system. This conveys the correctness of the recognisers' results when the evaluation data are given to the system:

$$WRR = \frac{WCR}{TWA} \times 100$$

In which *WCR* is the number of words correctly recognised and *TWA* is the total words attempted.

2. *Normalised Root Mean Square Error (NRMSE):* It is used to measure the accuracy of the system. NRMSE is usually measured in computational neurosciences in order to show how well a system learns a model. Here, it is based on the calculation of the absolute distance between the ideal results (i.e. zero and one as the min and max of the sigmoid activation function) and the actual results produced by the ASR system during the evaluation procedures. This parameter shows how close the ASR results are to the ideal ones in practice. Lower NRMSE percentage values show that the ASR is more accurate. NRMSE is simply defined as:

$$NRMSE(\%) = \frac{RMSE}{Max_{IdealOutput} - Min_{IdealOutput}}$$

where RMSE is calculated as:

$$RMSE = \sqrt{\frac{\sum_{j=1}^{m}\sum_{i=1}^{n}(IdealOutput_i - ANNOutput_i)^2}{n \times m}}$$

In which m is the number of evaluation samples and n is the vocabulary size. The parameters $Max_{IdealOutput}$ and $Min_{IdealOutput}$ were set as maximum and minimum of Sigmoid activation function.

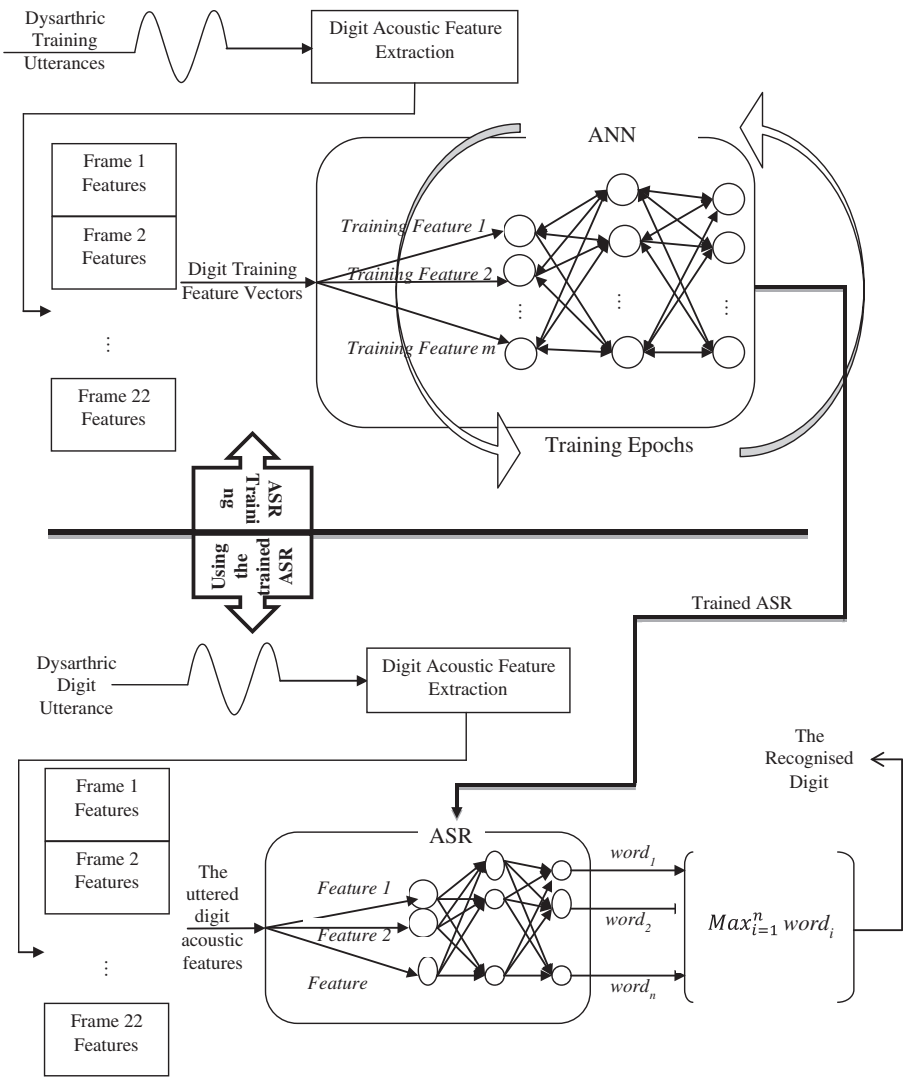

Fig. 1. Providing the ANN-based ASR system for users with dysarthria.

4. Experiments and results

In this section, we explain the two sets of experiments carried out in achieving the objectives of this study and present the evaluation results. They are described in detail as follows:

4.1. Experiment 1: identifying the best-performing set of MFCC parameters

The first set of experiments was conducted to identify the best-performing set of MFCC features for representing dysarthric acoustic signals in order to provide an ANN-based dysarthric ASR system. We trained 28 ANN-based speech recognisers (speaker-dependent) for seven dysarthric subjects. Four SD speech recognisers were provided for each subject, each of which was trained with a different set of MFCC parameters. These sets are:

(a) 12 MFCCs (i.e. mel cepstrum with 12 coefficients).
(b) 12 MFCCs first derivatives.
(c) 12 MFCCs second derivatives.
(d) Set *a* + Set *b* + Set *c*.

The database provided three different utterances of each digit per speaker. As training samples, we used one of the utterances together with the speech samples of three control speakers per digit. The second and third utterances were considered as ASR evaluation data and were employed to measure the aforementioned evaluation criteria.

For each dysarthric subject, the four MLPs were trained with the same tool and the same acoustic samples, but the extracted features were different. The tool was developed in our lab in Matlab and Visual Studio.Net. The trained MLPs had 60 hidden neurons, the activation function was Sigmoid, and 5000 training epochs were performed. It is important to note that the number of hidden neurons, training epochs, and activation functions was chosen by trial and error. For the MLPs trained by acoustic feature set *d*, each frame included mel cepstrum with 12 coeffs (set *a*), their first derivatives (set *b*), and their second derivatives (set *c*), (i.e. each frame was composed of 36 MFCC features). They were trained with the same parameters, but 130 hidden neurons were chosen, because the number of input neurons was bigger. Table 2 shows the results of these experiments.

4.2. Experiment 2: speaker-independent dysarthric ANN-based ASR

The evaluation of the proposed isolated-word SI dysarthric ASR model is discussed in this section. The speech materials of the 13 speakers with dysarthria described in Table 1 together with those of nine speakers without dysarthria, (i.e. control speakers) were

Table 2
Speaker-dependent ASR experimental results.

SD ASR no.	Participant[A]	Gender	Age	Speech intelligibility (%)	Acoustic feature set[B]	WRR (%)	NRMSE (%)
1	M04	Male	18	2	*a*	40.00	30.54
2					*b*	25.00	37.37
3					*c*	30.00	34.23
4					*d*	40.00	31.72
5	F03	Female	51	6	*a*	57.89	27.81
6					*b*	36.84	30.11
7					*c*	42.10	26.08
8					*d*	47.36	27.71
9	M07	Male	58	28	*a*	70.00	20.39
10					*b*	42.10	27.85
11					*c*	70.00	22.25
12					*d*	65.00	20.98
13	F02	Female	30	29	*a*	68.42	24.84
14					*b*	36.84	28.65
15					*c*	42.10	30.99
16					*d*	52.63	25.81
17	M06	Male	18	39	*a*	95.00	14.38
18					*b*	50.00	27.25
19					*c*	55.00	23.91
21					*d*	85.00	19.16
21	M05	Male	21	58	*a*	85.00	15.39
22					*b*	60.00	23.21
23					*c*	75.00	20.56
24					*d*	70.00	18.42
25	M09	Male	18	86	*a*	80.00	18.00
26					*b*	75.00	23.84
27					*c*	65.00	27.41
28					*d*	70.00	20.18

[A] Male control speakers were CM01, CM04, CM06, and female control speakers were CF02, CF03, CF05.
[B] The acoustic feature sets are (*a*) 12 MFCCs, (*b*) 12 MFCCs first derivatives, (*c*) 12 MFCCs second derivatives, and (*d*) *a* + *b* + *c*.

Table 3
Results of the speaker-independent dysarthric ASR system trained with acoustic feature set *a*.

Evaluation dataset	Subject	Number of evaluation samples	Correct classifications	Incorrect classification	WRR (%)	NRMSE (%)
High dysarthric severity	M07	77 Digits	44 Digits	33 Digits	57.14	24.53
Moderate dysarthric severity	M05	95 Digits	71 Digits	24 Digits	74.73	21.34
Mild dysarthric severity	F05	89 Digits	60 Digits	29 Digits	67.41	22.77
All dysarthric testing data	M07, M05, F05, plus silence	272 Samples	186 Samples	86 Samples	68.38	22.34
Testing data for speakers without disability	CF05, CM05	52 Digits	51 Digits	1 Digit	98.07	8.12

Table 4
Results of the speaker-independent dysarthric ASR system trained with acoustic feature set *d*.

Evaluation dataset	Subject	Number of evaluation samples	Correct classifications	Incorrect classification	WRR (%)	NRMSE (%)
High dysarthric severity	M07	77 Digits	38 Digits	39 Digits	49.35	27.77
Moderate dysarthric severity	M05	95 Digits	71 Digits	24 Digits	74.73	19.32
Mild dysarthric severity	F05	89 Digits	48 Digits	41 Digits	53.93	27.03
All dysarthric testing data	M07, M05, F05, plus silence	272 Samples	168 Samples	104 Samples	61.76	24.24
Testing data for speakers without disability	CF05, CM05	52 Digits	47 Digits	5 Digits	90.38	12.01

considered for training; the speech materials of the other three subjects with dysarthria were used for evaluation. The evaluation subjects with dysarthria were M07 for High Dysarthric Severity evaluation (28% speech intelligibility), M05 for Moderate Dysarthric Severity evaluation (speech intelligibility 58%), and F05 for Mild Dysarthric Severity evaluation (speech intelligibility 95%). The data of CF05 and CM05 were also considered during the evaluation process in order to study the behaviour of the proposed SI ASR model when it was used to process unforeseen normal speech data. It is pertinent to note that no speech data and sample of the evaluation (test) subjects (both dysarthric and non-disabled subjects) were provided to the recognisers during the training procedures.

Two SI speech recognisers were provided for subjects with dysarthria. For the first system, the mel cepstrum with 12 coefficients, (i.e. acoustic feature set *a*) was extracted as acoustic features, but

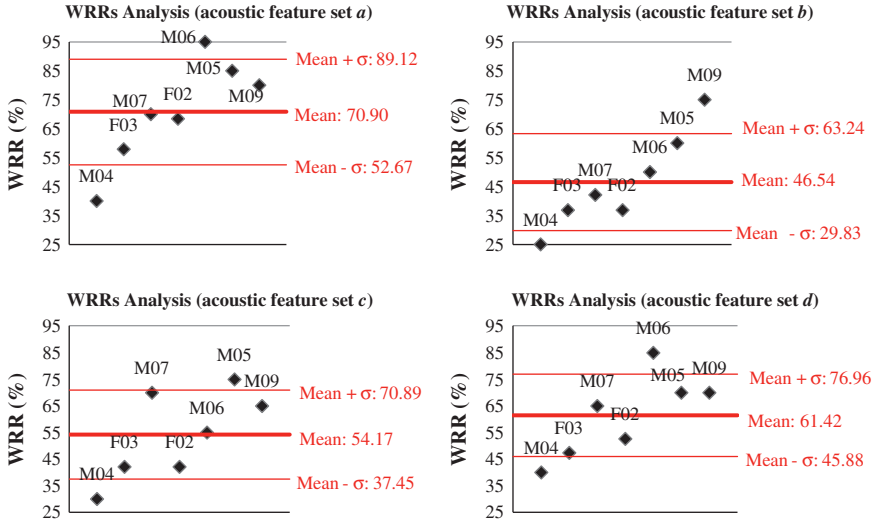

Fig. 2. Statistical analysis of WRRs (speaker-dependent experiments).

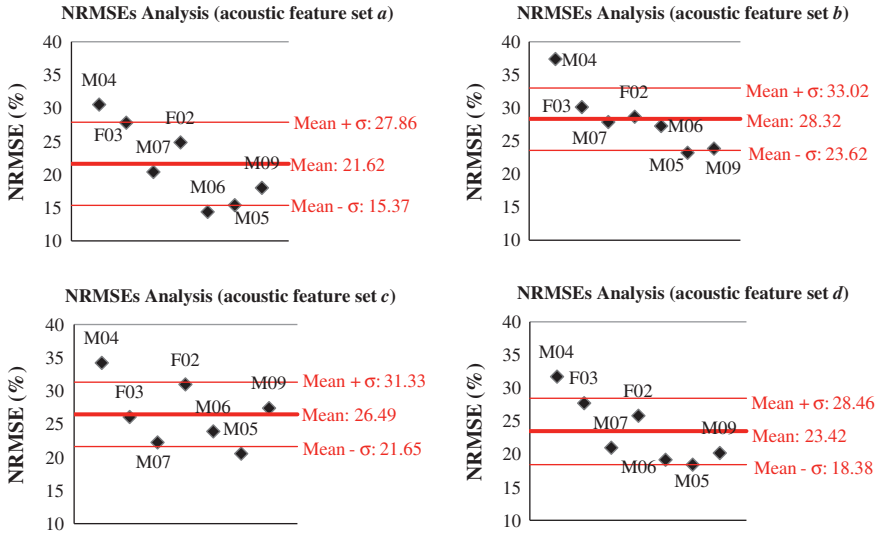

Fig. 3. Statistical analysis of NRMSEs (speaker-dependent experiments).

acoustic feature set *d* was considered for the second system. These sets of MFCC parameters were chosen, because we noted that these parameters provided better performance than the other two in SD experiments, as shown in Table 2.

Both SI recognisers consisted of 11 output neurons (one additional neuron to recognise silence) and one hidden layer, and training cycles was 5000. However, the number of input neurons was 264 for the first ASR system and 792 for the second ASR system (36 MFCC features per frame × 22 frames = 792 MFCC features). The chosen numbers of hidden neurons (h) were set as [34]:

$$h = ((c) \times (i)) + n$$

where i is the number of input neurons, $c = 2/3$ is a constant, and n is the vocabulary size. The training MSE of the first MLP was 0.03566 and the other was 0.00726.

As explained above, the performance of the speech recognisers presented in this section was evaluated by measuring the WRR as the rate of correct identification of digits by the recognisers, and the NRMSE was computed to measure their accuracy. The ideal result is "1" for a correctly identified digit and "0" otherwise. Therefore, any other values produced by the output neurons were regarded as distance errors and were considered while calculating NRMSE.

Tables 3 and 4 illustrate the results obtained by applying the testing datasets to the proposed SI speech recognisers accordingly. The results of each experiment with its associated set of MFCC parameters are presented in a separate table. The evaluation parameters were measured for each dysarthric severity level, (i.e.

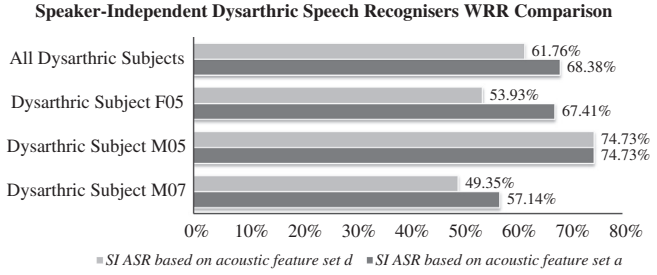

Fig. 4. Word recognition rate comparison of the two speaker-independent ASR systems.

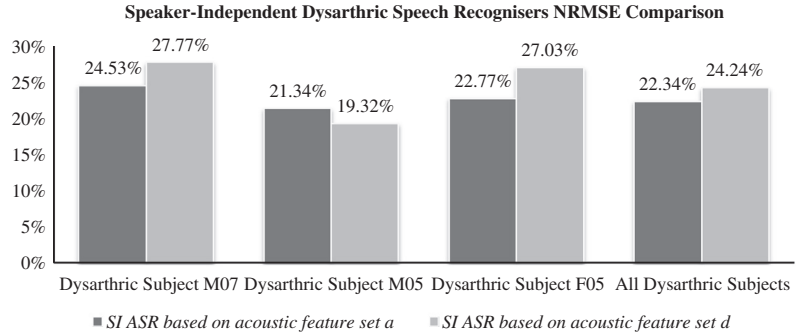

Fig. 5. NRMSE Comparison of the two speaker-independent ASR systems (lower is better).

for each dysarthric subject) separately. It should be noted that the "All Dysarthric Testing Data" dataset includes all of the speech samples obtained from the three dysarthric evaluation subjects, in addition to silence samples.

5. Discussion

In order to achieve the first objective of this paper, we provided 28 speaker-dependent ANN-based ASR systems over seven dysarthric subjects and trained them with different sets of MFCC parameters. Figs. 2 and 3 depict statistical analysis of the experiments employed to identify the best-performing set of MFCC parameters. It is shown that the WRRs and NRMSEs followed a Normal distribution because the minimum of 71.42% of the observations fell between *Mean* ± *SD* and 100% between *Mean* ± *2SD*. The speech recognisers that were trained with the 12 MFCCs as dysarthric speech features (set *a*) provided the highest mean WRR. In addition to the best recognition rates, they also produced the smallest mean NRMSE, which shows that the results from these experiments are more accurate.

Likewise, the results from the speech recognisers that were fed with mel cepstrum with 12 coefficients inclusive of their first and second derivatives are acceptable, although they are not as good as the results of speech recognisers fed with set *a* acoustic features. Nonetheless, it is shown that the first or second derivatives of mel cepstrum coeffs are not good dysarthric speech representations, because the speech recognisers trained with these parameters (sets *b* and *c*) performed poorly in terms of accuracy and WRR.

The second objective of this study is to investigate the application of ANNs in a fixed-length isolated-word speaker-independent dysarthric ASR model. Two ANN-based speech recognisers were provided using the acoustic materials collected from 13 dysarthric subjects in addition to the data of nine control speakers. Acoustic feature set *a* was used to represent dysarthric speech as input to the first SI ASR system; the acoustic features of the second ASR system were a combination of all the MFCC-based parameters, (i.e. acoustic feature set *d*). Figs. 4 and 5 compare the WRR and NRMSE of these two SI ASR systems respectively, for each of the evaluation datasets. The SI ASR system based on acoustic feature set *a* had consistently produced better results than the ASR system based on set *d*. There was only one occasion the ASR system trained with set *d* delivered lower NRMSE than the ASR system based on set *a*, namely the evaluation of dataset for subject M05.

Although the results of the SI ASR system, trained with set *d*, are quite close to those of set *a*, it is not recommended to consider all the MFCC parameters together for ANN-based, speaker-independent dysarthric ASR systems. The reason being that it needs a huge ANN to learn the acoustic features; it may not be practical and effective in terms of performance when the system is deployed in low-capacity devices. In our experiments, the SI ASR system trained with acoustic feature set *d* used an ANN with 1338 neurons and 430 K synaptic weights; training the system by using such a big ANN requires a considerable amount of computational resources. On the other hand, when the ASR system was trained with 12 MFCCs, it used an ANN with only 462 neurons and about 51 K synaptic weights. This means ANN trained with set *a* is much smaller than the ANN trained with set *d*. In addition, the ASR system which used mel cepstrum with 12 coefficients produced better results. Judging from the better results of the ASR system trained with acoustic feature set *a*, it can be concluded that ANNs can classify dysarthric speech more accurately when dysarthric acoustic features are presented as mel cepstrum with 12 coefficients.

6. Conclusions

In this paper we studied the application of ANNs in an SI ASR model for individuals with dysarthria. In addition, several SD ANN-based speech recognisers for users with dysarthria were provided, and the results were compared in detail. The purpose is to investigate and to ascertain the best MFCC-based feature set that can represent dysarthric acoustic features; the representation is then used by an ANN-based SI ASR system designed for individuals with dysarthria. The performance of the proposed ASR models was measured in terms of word recognition rate and accuracy of evaluation. Speech samples of the subjects with speech disabilities were evaluated by the proposed SI ASR systems for each dysarthric severity level separately. The speech data of the evaluation subjects were not included for the training of the SI speech recognisers. This exclusion of speech data of the evaluation subjects allows the generalisability of the proposed models to be evaluated.

The results show that mel cepstrum with 12 coefficients can be selected as the best set of MFCC acoustic features in order to train an ANN-based ASR system for speakers with dysarthria. The WRR of the dysarthric SI ASR model, trained with mel cepstrum including 12 coefficients achieved an average of 68.38%. It also produced 98.07% WRR for the speech of unanticipated speakers without speech disabilities. The highest WRR of speaker-dependent ASR models was 95%.

Acknowledgement

This work is funded by the University of Malaya under High Impact Research Grant (Grant No.: UM.C/HIR/MOHE/FCSIT/05).

References

[1] K. Rosen, S. Yampolsky, Automatic speech recognition and a review of its functioning with dysarthric speech, Augment. Altern. Comm. 16 (2000) 48–60.
[2] P.D. Polur, G.E. Miller, Effect of high-frequency spectral components in computer recognition of dysarthric speech based on a mel-cepstral stochastic model, J. Rehabil. Res. Dev. 42 (2005) 363–371.

[3] S.-A. Selouani, M.S. Yakoub, D. O'Shaughnessy, Alternative speech communication system for persons with severe speech disorders, EURASIP J. Adv. Signal Process. 2009 (2009) 1–12.
[4] S.O.C. Morales, S.J. Cox, Modelling errors in automatic speech recognition for dysarthric speakers, EURASIP J. Adv. Signal Process. 2009 (2009) 1–14.
[5] S.A. Borrie, M.J. McAuliffe, J.M. Liss, Perceptual learning of dysarthric speech: a review of experimental studies, J. Speech Lang. Hear. Res. 55 (2012) 290–305.
[6] P. Kitzing, A. Maier, V.L. Ahlander, Automatic speech recognition (ASR) and its use as a tool for assessment or therapy of voice, speech, and language disorders, Logop. Phoniatr. Voco. 34 (2009) 91–96.
[7] P.C. Doyle, H.A. Leeper, A.L. Kotler, N. Thomas-Stonell, C. Oneill, M.C. Dylke, K. Rolls, Dysarthric speech: a comparison of computerized speech recognition and listener intelligibility, J. Rehabil. Res. Dev. 34 (1997) 309–316.
[8] L. Ferrier, H. Shane, H. Ballard, T. Carpenter, A. Benoit, Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition, Augment. Altern. Comm. 11 (1995) 165–175.
[9] V. Young, A. Mihailidis, Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: a literature review, Assist. Technol. 22 (2010) 99–112.
[10] M.S. Hawley, P. Enderby, P. Green, S. Cunningham, S. Brownsell, J. Carmichael, M. Parker, A. Hatzis, P. O'Neill, R. Palmer, A speech-controlled environmental control system for people with severe dysarthria, Med. Eng. Phys. 29 (2007) 586–593.
[11] S.K. Fager, D.R. Beukelman, T. Jakobs, J.-P. Hosom, Evaluation of a speech recognition prototype for speakers with moderate and severe dysarthria: a preliminary report, Augment. Altern. Comm. 26 (2010) 267–277.
[12] K. Hux, J. Rankin-Erickson, N. Manasse, E. Lauritzen, Accuracy of three speech recognition systems: case study of dysarthric speech, Augment. Altern. Comm. 16 (2000) 186–196.
[13] E. Sanders, M. Ruiter, L. Beijer, H. Strik, Automatic recognition of Dutch dysarthric speech: a pilot study, in: Proceedings of the 7th International Conference on Spoken Language Processing, Denver, CO, USA, 2002, pp. 661–664.
[14] N. Talbot, Improving the speech recognition in the ENABL project, KTH TMH-QPSR 41 (2000) 31–38.
[15] P.D. Polur, G.E. Miller, Investigation of an HMM/ANN hybrid structure in pattern recognition application using cepstral analysis of dysarthric (distorted) speech signals, Med. Eng. Phys. 28 (2006) 741–748.
[16] M. Hasegawa-Johnson, J. Gunderson, A. Perlman, T. Huang, HMM-based and SVM-based recognition of the speech of talkers with spastic dysarthria, in: Proceedings of the 2006 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2006, pp. 1060–1063.
[17] H.A. Bourlard, N. Morgan, Connectionist Speech Recognition: A Hybrid Approach, Kluwer Academic Publishers, Norwell, MA, USA, 1994.
[18] P. Green, J. Carmichael, A. Hatzis, P. Enderby, M. Hawley, M. Parker, Automatic speech recognition with sparse training data for dysarthric speakers, in: Proceedings of the 8th European Conference on Speech Communication and Technology, Geneva, Switzerland, 2003, pp. 1189–1192.
[19] H.V. Sharma, M. Hasegawa-Johnson, State-transition interpolation and MAP adaptation for HMM-based dysarthric speech recognition, in: Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies, Association for Computational Linguistics, Los Angeles, CA, 2010, pp. 72–79.
[20] D. Jurafsky, J.H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, second ed., Pearson Prentice Hall, 2008.
[21] M. Wiśniewski, W. Kuniszyk-Jóźkowiak, E. Smołka, W. Suszyński, Automatic detection of disorders in a continuous speech with the hidden Markov models approach, in: M. Kurzynski, E. Puchala, M. Wozniak, A. Zolnierek (Eds.), Computer Recognition Systems 2, Springer, Berlin/Heidelberg, 2007, pp. 445–453.
[22] J.I. Godino-Llorente, P. Gomez-Vilda, Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors, IEEE Trans. Biomed. Eng. 51 (2004) 380–384.
[23] P.D. Polur, G.E. Miller, Experiments with fast Fourier transform, linear predictive and cepstral coefficients in dysarthric speech recognition algorithms using hidden Markov model, IEEE Trans. Neural Syst. Rehabil. Eng. 13 (2005) 558–561.
[24] J.R. Deller, D. Hsu, L.J. Ferrier, On the use of hidden Markov modeling for recognition of dysarthric speech, Comput. Meth. Prog. Biomed. 35 (1991) 125–139.
[25] H.V. Sharma, M. Hasegawa-Johnson, Universal access: speech recognition for talkers with spastic dysarthria, in: Proceedings of the 10th Annual Conference of the International Speech Communication Association 2009, Brighton, England, 2009, pp. 1447–1450.
[26] E. Trentin, M. Gori, A survey of hybrid ANN/HMM models for automatic speech recognition, Neurocomputing 37 (2001) 91–126.
[27] G. Dede, M.H. Sazli, Speech recognition with artificial neural networks, Digit. Signal Process. 20 (2010) 763–768.
[28] S.R. Shahamiri, S.S. Binti Salim, Real-time frequency-based noise-robust automatic speech recognition using multi-nets artificial neural networks: a multi-views multi-learners approach, Neurocomputing, in press, <http://www.sciencedirect.com/science/article/pii/S0925231213009661>.
[29] G. Jayaram, K. Abdelhamied, Experiments in dysarthric speech recognition using artificial neural networks, J. Rehabil. Res. Dev. 32 (1995) 162–169.
[30] F. Rudzicz, Using articulatory likelihoods in the recognition of dysarthric speech, Speech Commun. 54 (2012) 430–444.
[31] P. Raghavendra, E. Rosengren, S. Hunnicutt, An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems, Augment. Altern. Comm. 17 (2001) 265–275.
[32] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, S. Frame, Dysarthric speech database for universal access research, in: Proc. of the 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, 2008, pp. 1741–1744.
[33] R.D. Kent, Research on speech motor control and its disorders: a review and prospective, J. Commun. Disord. 33 (2000) 391–428.
[34] S.R. Shahamiri, W.M.N.W. Kadir, S. Ibrahim, S.Z.B. Hashim, An automated framework for software test oracle, Inf. Softw. Technol. 53 (2011) 774–788.