Mohammed *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2020) 2020:1
https://doi.org/10.1186/s13636-019-0169-5

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Improving dysarthric speech recognition using empirical mode decomposition and convolutional neural network

Mohammed Sidi Yakoub[1*], Sid-ahmed Selouani[1], Brahim-Fares Zaidi[2] and Asma Bouchair[3]

Abstract

In this paper, we use empirical mode decomposition and Hurst-based mode selection (EMDH) along with deep learning architecture using a convolutional neural network (CNN) to improve the recognition of dysarthric speech. The EMDH speech enhancement technique is used as a preprocessing step to improve the quality of dysarthric speech. Then, the Mel-frequency cepstral coefficients are extracted from the speech processed by EMDH to be used as input features to a CNN-based recognizer. The effectiveness of the proposed EMDH-CNN approach is demonstrated by the results obtained on the Nemours corpus of dysarthric speech. Compared to baseline systems that use Hidden Markov with Gaussian Mixture Models (HMM-GMMs) and a CNN without an enhancement module, the EMDH-CNN system increases the overall accuracy by 20.72% and 9.95%, respectively, using a *k*-fold cross-validation experimental setup.

Keywords: Dysarthria, Empirical mode decomposition, Hurst mode selection, Convolutional neural network

1 Introduction

Dysarthria is a motor speech disorder resulting from diverse impairments afflicting the control and execution of speech movements. A lack of coordination or a weakness of the muscles required for speech is then noticed. The main consequence of dysarthria is the degradation of speech intelligibility caused by poor articulation of consonants and, in the most severe cases, by the distortion of vowels [1]. Due to these important variations, a speech recognition system specifically tailored to dysarthric speakers would be more suitable than a generic system [2–4].

1.1 Related work

Most conventional dysarthric speech recognition systems are based on structured approaches. For instance, HMM-GMM-based approaches [5] use Hidden Markov Models (HMMs) to model the sequential structure of the speech signal and Gaussian Mixture Models (GMMs) to model the distribution of the spectral representation of a waveform. However, HMM-GMM-based systems require a large amount of data to be trained, which is not efficient in the case of dysarthric speech where the corpora used for training are always small [6]. Therefore, these approaches cannot be applied with ease in the context of dysarthric speech.

In recent years, the field of pathological speech processing has seen major breakthroughs in effective alternatives to HMM-GMM that can better recognize dysarthric speech thanks to the development of deep neural network (DNN) architectures. For instance, the work by Kim et al. [4] adopts convolutional long short-term memory recurrent neural networks to model dysarthric speech in a speaker-independent way, taking advantage of convolutional neural networks to extract effective local features. The approach in Vachhani et al. [7] uses healthy speech augmented with simulated dysarthric speech to train a DNN-HMM-based automatic speech recognition. Bhat et al. [8] proposed using a time-delay neural network-based denoising autoencoder to enhance dysarthric speech features before performing DNN-HMM–based recognition.

*Correspondence: mohammed.sidi.yakoub@umoncton.ca
[1]Moncton University, UMCS Shippagan, 218 Boul. J.D. Gauthier, Shippagan, NB, E8S 1P6, Canada
Full list of author information is available at the end of the article

Liu et al. [9] investigated the use of pitch features for disordered speech recognition. They attempt to use gated neural network and Bayesian gated neural network models to explore the robust integration of pitch features to improve the recognition performance for disordered speech. The authors in [10] used a multilayer perceptron (MLP) artificial neural network for diagnostic testing of speech affected by Parkinson's disease. He used a relatively large dataset of speech sample features compiled from Parkinson's disease patients and healthy individuals to train the MLP system to accurately classify between these two speech types. The dysarthric speech recognition system proposed by Hu et al. [11] is based on a gated neural network (GNN) and allows robust integration of acoustic features with visual features as well as, optionally, prosody features based on pitch. They designed two systems for English and Cantonese languages where a speaker-independent GNN acoustic model was trained for the mixed training set using a concatenated 86-dimension log filter bank plus pitch features.

The authors in [12] investigated the potential correlation between noisy speech and dysarthric speech in terms of intelligibility. They specifically examined whether there is a relationship between the ability of an individual to deal with noisy speech uttered in severely degraded environments and his/her ability to understand dysarthric speech. The metrics used revealed that listeners who were able to understand speech in noise also succeeded in understanding dysarthric speech. Although the origin of the degradation differs, we can reasonably consider that the speech in noise results from the degradation of the environment, while dysarthric speech results from the degradation of the source.

1.2 Motivation and objective of the study

In this paper, dysarthric speech is viewed as noisy speech, and therefore, we recommend the use of speech enhancement techniques to improve its quality and thus its intelligibility. A combination of a robust variant of empirical mode decomposition (EMD) with a convolutional neural network is proposed to perform an accurate phonemic recognition of dysarthric speech.

The motivation underlying the use of EMD is its ability to produce a time-frequency representation of non-stationary signals using an adaptive decomposition of the signals in the sum of oscillating components called intrinsic mode functions (IMFs). This decomposition has the appearance of a generalized Fourier analysis with variable amplitudes and frequencies. The frequency resolution of EMD for any point is uniformly defined by the stationary phase or local derivative of the phase, which makes this method appropriate for the effective extraction of low-frequency oscillations that are often observed in dysarthric speech.

The rest of this paper is organized as follows. Section 2 presents the Hurst exponent, the empirical mode decomposition, and the Hurst-based mode selection (EMDH) used to enhance the quality of dysarthric speech. The architecture of the proposed system and the HMM-GMM- and CNN-based recognizers are depicted in this section. Section 3 describes the experimental setup, the data in the Nemours corpus used in the experiments, and the obtained results. Section 4 concludes the paper and provides an overview of future work.

2 Proposed method

2.1 Hurst exponent estimation

The Hurst exponent H was named after the British hydrologist Harold Edwin Hurst, who studied the long-term storage capacity of reservoirs [13] and discovered the presence of long-term dependence in hydrology. Many years later, the mathematician Benoit Mandelbrot [14] generalized the Hurst exponent as a measure of the long-term memory of a time series, i.e., the tendency of a time series to strongly regress towards its mean or to cluster in one direction.

Since then, the Hurst exponent has been widely used as a measure of time dependence in various time series. The estimate of its value gives a hint on the persistence/autocorrelation in the related time series: if $0 < H < 0.5$, the series is antipersistent or has a negative autocorrelation; if $H = 0.5$, there is no autocorrelation in the series; and if $0.5 < H < 1$, the series is persistent and has a positive autocorrelation.

In the spectral domain, the Hurst exponent is related to the spectral characteristics of the time series [15]. For $H = 0.5$, the time series has equal intensity at different frequencies, and the power spectral density $S(f)$ of the signal is constant, e.g., white noise; on the other hand, if $0.5 < H < 1$, low frequencies are prominent, especially when H is closer to 1, i.e., $S(f)$ is inversely proportional to the frequency of the signal (pink noise).

Due to the characteristics above, the Hurst exponent estimation has been applied in different areas ranging from hydrology to speech processing. For instance, the authors in [16] composed speech vectors based on the Hurst exponent to perform speaker recognition, and the authors in [15] used the Hurst exponent to estimate noise and enhance speech signals.

Different Hurst exponent estimation methods exist, but this work uses discrete wavelet transforms (DWT) as in [15]. Each IMF is split into non-overlapping frames, and the steps below briefly describe how the discrete wavelet transform is used to estimate the Hurst exponent:

1 Apply the discrete wavelet transform:
A time domain signal is decomposed successively by the DWT into approximation $a_j(n)$ and detail $d_j(n)$

coefficients, where j is the decomposition scale and n is the coefficient index of each scale.

2 Estimate the variance σj^2:
Denoising techniques based on DWT exploit the detail coefficients to produce a smoother signal. Hence, the variance is estimated from those coefficients as:

$$\sigma j^2 = (1/N_j) \sum{}_n d_j(n)^2$$

where N_j is the number of available coefficients for each scale j.

3 Calculate the Hurst exponent:
The Hurst exponent is estimated as $H = (1 + \theta)/2$, where θ is the slope of the plot of

$$y_j = log_2(\sigma j^2)$$

versus j. The slope θ is obtained with a weighted linear regression.

2.2 EMDH for enhancing dysarthric speech

The principle of the EMDH speech enhancement technique consists of identifying and selecting the intrinsic mode functions (*IMFs*) that are less distorted and using them to reconstruct the enhanced speech signal [15].

First, EMDH starts with the decomposition of a dysarthric speech signal into a set of *IMFs*. Then, each *IMF* is windowed into non-overlapping short-time frames to determine the most distorted *IMF* by calculating the Hurst exponent for each *IMF* and using the wavelet-based method to generate Daubechies filters [15]. This frame-by-frame analysis is performed to avoid the incorrect selection of IMF_i due to the extreme variability and non-stationarity of the signal. Finally, the frames with Hurst exponent values greater than a given threshold H_{th} are dropped, and the remaining frames are used to reconstruct the enhanced dysarthric speech. The EMDH steps are as follows:

1 Use the EMD algorithm to decompose a dysarthric speech signal $x(t)$ into a finite number of IMF_s. If $\boldsymbol{M}$ IMF_s are extracted and the m^{th} is denoted as IMF_m, then

$$\left[x(t) = \sum_{m=1}^{M} IMF_m(t) + r(t) \right] \tag{1}$$

where $r(t)$ is the residual signal remaining after the decomposition.

2 Split each IMF_m into Q non-overlapping short-time frames. We obtain the following window IMF_m ($wIMF_m$):

$$wIMF_{m,q}(t) = \left\{ \begin{array}{ll} IMF_m(t + qT_d), & t \in [0, T_d] \\ 0, & \text{elsewhere} \end{array} \right\} \tag{2}$$

where $q \in \{0, \ldots, Q-1\}$ is the frame index and T_d is the fixed time duration of the frames.

3 Apply wavelet decomposition to all $wIMF_m$. Estimate their Hurst exponents and store them in a vector of Hurst values $H_q(m)$ with M components ($m \in \{1, \ldots, M\}$) calculated for each index q.

4 Determine for each frame the index N_q of the last $wIMF_m$ whose Hurst value is below a given threshold: $H_q(N_q) < H_{th}$.

5 Reconstruct the enhanced speech signal $\hat{x}(t)$ as

$$\hat{x}(t) = \sum_{q=0}^{Q-1} \hat{x}_q(t - qT_d) \tag{3}$$

and

$$\hat{x}_q(t) = \sum_{m=1}^{N_q} wIMF_{m,q}(t) \tag{4}$$

$q \in \{0, \ldots, Q-1\}$.

We set M to 10 and H_{th} to 0.90 in the experiments.

To show the effect of the EMDH technique, we conducted enhancement experiments on dysarthric speech using spectral subtraction, Wiener filtering, and EMDH techniques. The analysis of the obtained results showed that EMDH always gives the higher segmental signal-to-noise ratio (SNRseg). The perceptual evaluation of speech quality (PESQ) metric also illustrated the superiority of EMDH over the two other techniques. Therefore, we adopted EMDH as a preprocessing step in our proposed system. Table 1 gives the results of the enhancement experiment using a sentence from dysarthric speaker JF.

2.3 HMM-GMM system

HMMs model the sequential structure of the speech signal, and GMMs model the distribution of the spectral representation of the signal using different Gaussian distributions.

The HMM-GMM speech recognition system is built using HTK tools [5], where each phoneme is modeled by a 5-state HMM model with 2 non-emitting states (the first and fifth states) and a mixture of 2, 4, 8, or 16 Gaussian distributions. Mel-frequency cepstral coefficients (MFCCs), delta coefficients, and the cepstral pseudo-energy are calculated for all utterances and used as parameters to train and test the system.

Table 1 Results of enhancement techniques using a sentence from dysarthric speaker JF extracted from the Nemours corpus

	Speech enhancement technique		
Metrics	Spectral subtraction	Wiener filtering	EMDH
SNRseg	2.4146	11.1018	*15.8155*
PESQ	2.1160	3.2983	*4.3640*

Figure 1 depicts the training and test phases of this system.

We used this HMM-GMM system as a baseline for comparison with the EMDH-CNN system.

2.4 CNN system

Convolutional neural networks (CNNs) perform the mathematical operation called convolution instead of general matrix multiplication to produce filtered feature cards stacked on the top of each other. A CNN has the following components [17]:

1. Convolutional layer:
   This layer extracts characteristics from the input features, and its output feature maps are given by $C(x_{u,v}) = \sum_{i=-\frac{n}{2}}^{\frac{n}{2}} \sum_{i=-\frac{m}{2}}^{\frac{m}{2}} f_k(i,j)\, x_{u-i,v-j}$ where f_k is the filter with kernel size $n \times m$ applied to the input x, and $n \times m$ is the number of input connections to each CNN neuron.
2. Pooling layer:
   The pooling layer is a downsampling strategy applied to the output feature maps of the convolutional layer to reduce their number and make them more invariant to changes in scale and orientation. Two common pooling functions are used to reduce the number of sub-regions: average pooling and max pooling. We use the max pooling function. $M(x_i) = \max\{x_{i+k,i+l} \mid |k| \leq \frac{m}{2},\ |l| \leq \frac{m}{2},\ k, l \in N\}$ where x is the input and m is the size of the max filter.
3. ReLU (rectified linear unit):
   This activation function replaces all negative values in the feature map with zero. The main goal of ReLU is to introduce non-linearity into the CNN system because most of the data we would like the system to learn would be non-linear. ReLU is defined by the equation $R(x) = \max(0, x)$, where x is the input. Other functions are used, such as *tanh* or *sigmoid*, but ReLU has been shown to be more efficient.
4. Fully connected layer:
   Adding a fully connected layer to a CNN system is another good way to learn non-linear combinations of the features coming from the previous layer. However, the fully connected layer takes all the neurons from the previous layer and connects them to each neuron it has. The output of this layer is given by $F(x) = \sigma(W_{l \times k} \times x)$ where k is the size of the input x, l is the number of neurons in the fully connected layer, and σ is the activation function. This results in a matrix W.
5. Output layer:
   This layer is a hot vector representing the class of the given input vector x and has a dimensionality equal to the number of classes. In this work, we have 44 classes. The resulting class is represented by $C(x) = \{i \mid \exists i \,\forall j \neq i : x_j \leq x_i\}$
6. Softmax layer:
   The error is propagated back over a Softmax layer. For an input vector x of dimension N, the Softmax

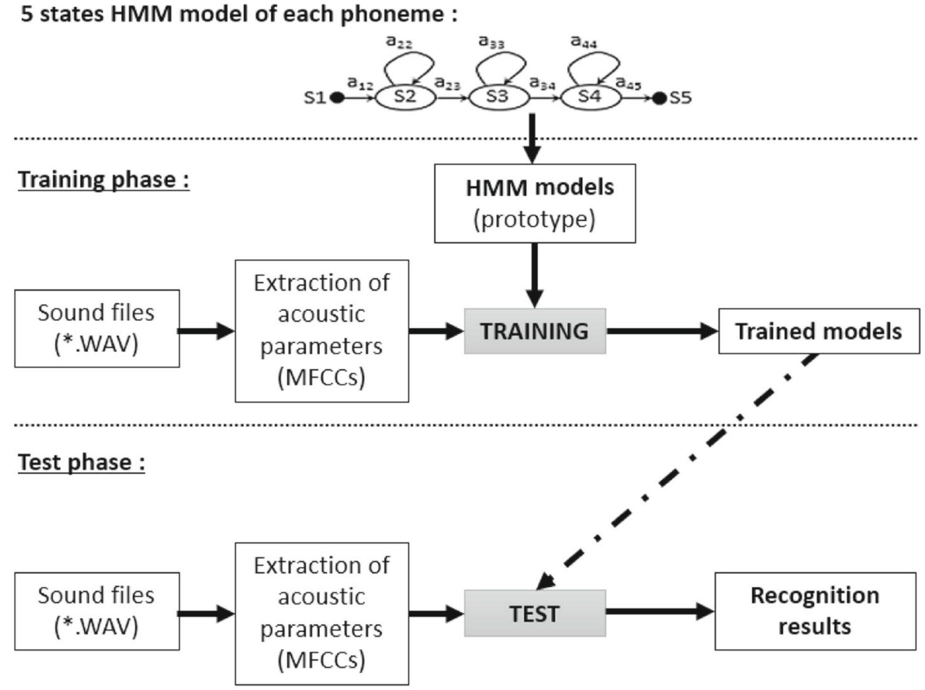

Fig. 1 HMM-GMM system architecture

layer calculates a mapping such that $S(x) : R^N \rightarrow [0,\ 1]^N$, and for each component $1 \leq j \leq N$, the output is calculated as follows: $S(x)_j = \frac{e^{x_j}}{\sum_{i=1}^{n} e^{x_i}}$

In general, a CNN consists of several iterations of this succession of layers where the output of one layer is the input to the next layer. One of the advantages of CNNs is that they are trained quickly. The architecture of our CNN system is depicted in Fig. 2.

We used this CNN system as a baseline for comparison with the EMDH-CNN system.

2.5 EMDH-CNN system for performing dysarthric speech recognition

The proposed EMDH-CNN system is depicted in Fig. 2. The dysarthric speech is first processed using the EMDH technique with the expected goal of enhancing its quality. In the second step, the Mel-frequency cepstral coefficients (MFCCs) are extracted from the enhanced speech and used as the input features to a deep learning-based system using a convolutional neural network (CNN). The CNN performs the recognition of 44 dysarthric phonemes.

The classification of dysarthric phonemes is performed by a CNN composed of a convolutional layer using a 39 ×

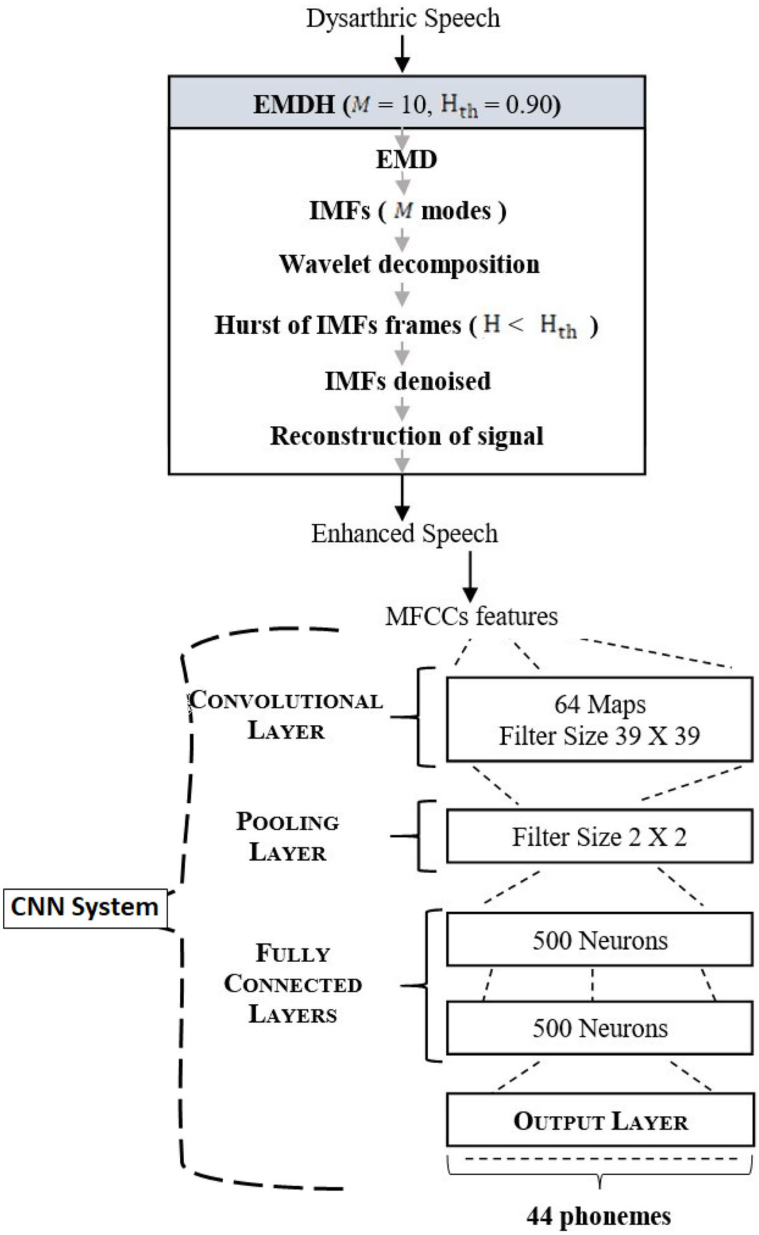

Fig. 2 EMDH-CNN system architecture

39 filter and 64 maps followed by a pooling layer with a 2 × 2 filter. A dropout of 0.2 is applied followed by a flattening layer and two fully connected layers of 500 neurons each. The last layer is the output layer with a size of 44, which corresponds to the number of phonemes to be recognized.

3 Experiments and results

The Nemours corpus was used throughout the experiments. This database is composed of 814 short non-sense sentences: 74 sentences each uttered by 11 American male speakers with different degrees of dysarthria [18]. This corpus was used to train the HMM-GMM and CNN baseline systems. The enhanced dysarthric speech utterances were obtained after processing the original Nemours corpus with the EMDH technique. This new corpus was used to train the EMDH-CNN system. Both the original corpus and the enhanced corpus were split into 70% and 30% to form training and test sets, respectively. We have also carried out experiments using 10-fold cross-validation to train and test speaker-dependent systems and global systems.

3.1 Experimental setup

The CNN- and HMM-GMM-based systems were used as baseline references. Their input features are MFCCs extracted from the dysarthric speech without any preprocessing. For the EMDH-CNN system, the MFCC features were extracted after preprocessing the corpus with the EMDH technique. Only the preprocessed MFCCs were used to train and test the EMDH-CNN system. Python programming language and Keras library [19] with TensorFlow were used to carry out the experiments.

The systems were tested using input vectors composed of 26 MFCCs and their first derivatives (13 + 13 deltas), which was found to be the best configuration for this application. Prior cross-validation experiments showed that the second derivatives do not improve the accuracy. Due to the difficulty of understanding dysarthric speech and the extreme variability observed between dysarthric speakers, the three systems were designed as speaker-dependent. The best parameters of the CNN and EMDH-CNN systems are depicted in Fig. 2.

The speaker-dependent HMM-GMM baseline system uses a monophone left-to-right model with Gaussian mixture output densities. Each phoneme is modeled by a 5-state HMM model where the first and last states are non-emitting. The best HMM-GMM system results were obtained with 4 Gaussian mixtures.

The use of a speaker-dependent monophone model is justified by the lack of exhaustive training data. Indeed, in the context of dysarthria, there is a huge variability in the pronunciation of a single phoneme, and therefore, a large amount of training data is required to train a robust model. Numerous studies, such as one conducted to adapt dysarthric speech, suggest using speaker-dependent monophone models as a baseline system [20].

3.2 Results and discussion

The accuracy of each system by speaker and the global accuracy of the three systems are shown in Tables 2 and 3, respectively. Table 2 clearly shows that the EMDH technique improves the accuracy of the CNN system for almost all speakers by an average of 2.07% and 4.90% using 70/30 split and 10-fold, respectively. The same conclusion can be drawn for the overall accuracy, which is improved by 2.63% and 9.95% using 70/30 split and 10-fold, respectively, as shown in Table 3.

The low performance obtained by the HMM-GMM system was predictable because of the extreme variability of dysarthric speech and the relatively small amount of data available for training. In this context, the assumption made by the GMM, that the input occurrences conform to a Gaussian distribution, is probably not valid. It is important to mention that dysarthric speech is characterized by phonetic articulatory errors and other diverse and unfavorable artifacts, such as stuttering, accidental pauses, and involuntary breathing.

The obtained results show that the CNN has a good ability to extract latent features of dysarthric speech and can be trained more easily with a limited amount of data.

4 Conclusion and future work

In this paper, a new approach has been proposed to improve dysarthric speech recognition. We used empirical mode decomposition and Hurst-based mode selection to enhance dysarthric speech as a preprocessing step. The robust dysarthric speech recognition system was designed by coupling the EMDH-based enhancement process with a convolutional neural network. The obtained results show that the EMDH-CNN system performed better than

Table 2 Speaker accuracies of the HMM-GMM, CNN, and EMDH-CNN dysarthric speech recognizers

	Systems			
Speaker	HMM-GMM (%)	CNN (%)	EMDH-CNN (70/30) (%)	EMDH-CNN 10-fold (%)
BB	50.00	69.43	72.59	*74.54*
BK	31.39	45.54	47.10	*49.01*
BV	46.00	54.50	54.62	*59.67*
FB	49.31	80.28	80.89	*83.22*
JF	45.14	58.69	61.57	*63.47*
LL	46.00	60.43	62.31	*69.91*
MH	34.67	73.00	75.37	*77.93*
RK	45.33	47.17	50.95	*53.73*
RL	42.00	67.14	67.73	*69.60*
SC	42.67	51.64	55.41	*55.84*

Table 3 Global accuracies of the HMM-GMM-, CNN-, and EMDH-CNN-based dysarthric speech recognition systems

	Systems			
All speakers	HMM-GMM	CNN	EMDH-CNN (70/30)	EMDH-CNN 10-fold
Accuracy (%)	44.13%	54.91%	57.54%	*64.86%*

the HMM-GMM and CNN baseline systems. While the results obtained are encouraging, it is still challenging to find accurate models because of the extreme variability and complexity of dysarthric speech. Future work will investigate the use of other robust acoustic features and a long-term, memory-based deep learning architecture.

Abbreviations

CNN: Convolutional neural network; DNN: Deep neural network; EMD: Empirical mode decomposition; EMDH: Empirical mode decomposition and Hurst-based mode selection; HMM-GMMs: Hidden Markov with Gaussian Mixture Models; IMFs: Intrinsic mode functions; MFCCs: Mel-frequency cepstral coefficients

Acknowledgments

Not applicable.

Authors' contributions

The main contributions are from SM and SS. BA prepared the comparison between the three speech enhancement techniques, and ZB built the HMM-GMM system. SS checked the entire article and polished the article structure. All authors read and approved the final manuscript.

Authors' information

The first two authors have been active in the field for two decades.

Funding

This work received financial support from the Natural Sciences and Engineering Research Council of Canada (1510-2018).

Availability of data and materials

No data were generated during this study but are available from [18]. The code can be provided to reproduce the results.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Moncton University, UMCS Shippagan, 218 Boul. J.D. Gauthier, Shippagan, NB, E8S 1P6, Canada. [2]LSCSP Research Laboratory, USTHB University, BP 32 Bab Ezzouar, 16111, Algiers, Algeria. [3]LCPTS Research Laboratory, USTHB University, BP 32 Bab Ezzouar, 16111, Algiers, Algeria.

Received: 19 August 2019 Accepted: 19 December 2019
Published online: 13 January 2020

References

1. P. Enderby, in *Handbook of Clinical Neurology (110 ed.)* Disorders of communication: Dysarthria (Elsevier B. V., 2013), pp. 273–281. https://www.sciencedirect.com/science/article/pii/B9780444529015000228. https://doi.org/10.1016/B978-0-444-52901-5.00022-8
2. 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**(8), 741–748 (2006)
3. M. Hasegawa-Johnson, J. Gunderson, A. Perlman, T. Huang, in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Toulouse*. Hmm-Based and Svm-Based Recognition of the Speech of Talkers With Spastic Dysarthria, (2006), pp. III-III. https://ieeexplore.ieee.org/abstract/document/1660840. https://doi.org/10.1109/ICASSP.2006.1660840
4. M. J. Kim, B. Cao, K. An, J. Wang, in *Interspeech*. Dysarthric speech recognition using convolutional lstm neural network, (2018), pp. 2948–2952. https://www.researchgate.net/publication/327350843_Dysarthric_Speech_Recognition_Using_Convolutional_LSTM_Neural_Network
5. S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, et al., The htk book (for htk version. 3.3), Cambridge University Engineering Department, 2005 (2006). http://htk.eng.cam.ac.uk/docs/docs.shtml
6. S. Oue, R. Marxer, F. Rudzicz, in *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*. Automatic dysfluency detection in dysarthric speech using deep belief networks, (2015), pp. 60–64. https://www.aclweb.org/anthology/W15-5111/
7. B. Vachhani, C. Bhat, S. K. Kopparapu, in *Interspeech*. Data augmentation using healthy speech for dysarthric speech recognition, (2018), pp. 471–475. https://www.iscaspeech.org/archive/Interspeech_2018/pdfs/1751.pdf. https://www.semanticscholar.org/paper/Data-Augmentation-Using-Healthy-Speech-for-Speech-Vachhani-Bhat/e98ea9dc73bf87e5509e987addf56b7006593ad7
8. C. Bhat, B. Das, B. Vachhani, S. K. Kopparapu, in *Interspeech*. Dysarthric speech recognition using time-delay neural network based denoising autoencoder, (2018), pp. 451–455. https://www.iscaspeech.org/archive/Interspeech_2018/pdfs/1754.pdf. https://www.researchgate.net/publication/327389525_Dysarthric_Speech_Recognition_Using_Timedelay_Neural_Network_Based_Denoising_Autoencoder
9. S. Liu, S. Hu, X. Liu, H. Meng, *On the use of pitch features for disordered speech recognition*, (2019), pp. 4130–4134. https://www.iscaspeech.org/archive/Interspeech_2019/pdfs/2609.pdf. https://www.semanticscholar.org/paper/On-the-Use-of-Pitch-Features-for-Disordered-Speech-Liu-Hu/26cd586e2e704cc46099e9af18c56b3d7419ec54
10. J. Wu, Application of artificial neural network on speech signal features for Parkinson's disease classification (2019). http://csusm-dspace.calstate.edu/handle/10211.3/209929
11. S. Hu, S. Liu, H. F. Chang, M. Geng, J. Chen, T. K. H. Chung, J. Yu, K. H. Wong, X. Liu, H. Meng, *The cuhk dysarthric speech recognition systems for English and Cantonese*, (2019), pp. 3669–3670. https://www.iscaspeech.org/archive/Interspeech_2019/pdfs/8047.pdf. https://www.semanticscholar.org/paper/The-CUHK-Dysarthric-Speech-Recognition-Systems-for-Hu-Liu/21f55376e76a0602525bdfbe54c00ff97226c30a
12. S. A. Borrie, M. Baese-Berk, K. Van Engen, T. Bent, A relationship between processing speech in noise and dysarthric speech. J. Acoust. Soc. Am. **141**(6), 4660–4667 (2017)
13. H. E. Hurst, Long-term storage capacity of reservoirs. Trans. Amer. Soc. Civil Eng. **116**, 770–799 (1951)
14. B. B. Mandelbrot, R. L. Hudson, *The (mis) Behaviour of Markets: a Fractal View of Risk, Ruin and Reward*. (Profile books, 2010). https://users.math.yale.edu/~bbm3/web_pdfs/misbehaviorprelude.pdf
15. L. Zao, R. Coelho, P. Flandrin, Speech enhancement with EMD and Hurst-based mode selection. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(5), 899–911 (2014)
16. R. Sant'Ana, R. Coelho, A. Alcaim, Text-independent speaker recognition based on the Hurst parameter and the multidimensional fractional brownian motion model. IEEE Trans. Audio Speech Lang. Process. **14**(3), 931–940 (2006)
17. P. Burkert, F. Trier, M. Z. Afzal, A. Dengel, M. Liwicki, Dexpression: deep convolutional neural network for expression recognition. arXiv preprint arXiv:1509.05371 (2015)
18. X. Menendez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, H. T. Bunnell, in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*. The Nemours database of dysarthric speech, (Philadelphia, 1996), pp. 1962–1965. https://ieeexplore.ieee.org/abstract/document/608020
19. F. Chollet, et al., Keras (2015). http://keras.io/. https://keras.io/getting-started/faq/#how-should-i-citekeras
20. K. T. Mengistu, F. Rudzicz, in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Adapting acoustic and lexical models to dysarthric speech, (Prague, 2011), pp. 4924–4927. https://ieeexplore.ieee.org/abstract/document/5947460. https://doi.org/10.1109/ICASSP.2011.5947460