Neural Comput & Applic (2012) 21:935–943
DOI 10.1007/s00521-011-0563-0

ORIGINAL ARTICLE

A novel neural-based model for acoustic-articulatory inversion mapping

**Hossein Behbood · Seyyed Ali Seyyedsalehi ·
Hamid Reza Tohidypour · Mojtaba Najafi ·
Shahriar Gharibzadeh**

Received: 15 June 2010 / Accepted: 7 February 2011 / Published online: 15 March 2011

Abstract In this paper, a new bidirectional neural network for better acoustic-articulatory inversion mapping is proposed. The model is motivated by the parallel structure of human brain, processing information by having forward–reverse connections. In other words, there would be a feedback from articulatory system to the acoustic signals emitted from that organ. Inspired by this mechanism, a new bidirectional model is developed to map speech representations to articulatory features. Formation of attractor dynamics in such bidirectional model is first carried out by training the reference speaker subspace as the continuous attractor. Then, it is used to recognize the other speaker's speech. In fact, the structure and training of this bidirectional model is designed in such a way that the network learns to denoise the signal step by step, using properties of attractors it has formed. In this work, the efficiency of a nonlinear feedforward network is compared to the same one with a bidirectional connection. The bidirectional model increases the accuracy up to approximately 3% (from 62.09 to 64.91%) in the phone recognition process.

Keywords Bidirectional neural networks (BNNs) · Feed-forward networks (FFNs) · Time delay neural networks (TDNNs) · MOCHA-TIMIT database · Acoustic-articulatory inversion mapping

1 Introduction

Automatic speech recognition (ASR) is strongly concentrated on the use of acoustic representations of speech as input data. Speech engineers believe that the acoustic signals of speech are the most important means of communication between humans. However, articulatory movements have meaningful correlation with acoustic energies emitted from the corresponding organ. Combined with acoustic representations, these data generate excellent results for an enhanced speech recognition, analysis, and synthesis [1]. Some efforts for using articulatory features in quest of having better speech recognition can be observed in [1, 2].

Before the existence of reliable and precise equipment, researchers used vocal-tract models or linguistic rules to produce suitable articulatory gesture representations [3]. Today, we have continuous smooth data measured with sophisticated equipment. However, the use of such equipment is impossible due to the high cost and complications. Therefore, efforts are focused on a method to estimate the articulatory features from the acoustic signal. Various estimation methods and their challenges have been a fundamental topic for research in this decade. Some efforts in the mapping of acoustic to articulatory features are a trajectory mixture density networks (TMDNs) model [4],

H. Behbood (✉) · S. A. Seyyedsalehi · H. R. Tohidypour · S. Gharibzadeh
Department of Biomedical Engineering,
Amirkabir University of Technology, Tehran, Iran
e-mail: hossein_1779@aut.ac.ir

S. A. Seyyedsalehi
e-mail: ssalehi@aut.ac.ir

H. R. Tohidypour
e-mail: hamidto86@aut.ac.ir

S. Gharibzadeh
e-mail: gharibzadeh@aut.ac.ir

M. Najafi
Azad University, South Tehran Branch, Tehran, Iran
e-mail: mnajafi.81@gmail.com

TMDNs with multiple mixtures [5], multitask learning perspective [6], modeling the uncertainty in recovering articulation from acoustics [7], Gaussian mixture model (GMM) [8], accurate recovery of articulatory positions from acoustics [9] and Hidden Markov model (HMM)-based inversion systems to recover articulatory movements from speech acoustics [10].

Recently, more attempts are reported using artificial neural networks (ANNs) [5, 11]. Moreover, the human perception system shows that it has a bidirectional structure [12]. Therefore, in this paper, we focus on a nonlinear mapping between the acoustic representations of speech and the articulatory features by the use of a new bidirectional neural network (BNN) model. This network is inspired by the parallel structure of human brain, processing information by having forward–inverse connections. In this method, the primary recognition is accomplished by reliable regions. Unreliable regions are corrected afterward. This action is iterated until recognition is completed, and the primary recognition is modified using the final recognition (attractors).

One of the prominent theories that explains these connections is Motor Articulatory Feedback theory [13]. According to this hypothesis, there is a biological feedback from the acoustic signals of speech to the human articulators. This feedback is controlled by the brain and makes the speech chain a closed-loop process. Motivated by this hypothesis, we aim to implement an adaptive neural network model that includes a successful inversion mapping process. The proposed model offers higher accuracy in comparison with a standard feed-forward network (FFN) model.

Several reports confirm the capability of time delay neural networks (TDNNs) in the phone recognition process [14]. In this study, we successfully employ a special structure of the TDNN for recognition processes. We use past and future inputs instead of using every input, individually. Briefly, Both FFN and BNN models apply to map acoustic representations of speech in the form of electromagnetic articulography features. The outputs of these models are passed to the TDNN model for better phone recognition. In all neural network structures proposed in this study, the resilient optimization algorithm is used to minimize error function.

2 Method and models

2.1 Database and pre-processing

The multichannel articulatory (MOCHA) database is a corpus of 460 TIMIT sentences of 40 different speakers [15]. This database includes acoustic signals, laryngograph (LAR), electropalatograph (EPG), and electromagnetic articulograph (EMA). Acoustic signals are recorded with a sampling frequency of 16,000 Hz. EMA sensors are connected to the upper and lower lips, lower incisor (jaw), tongue tip (5–10 mm from the tip), tongue blade (approximately 2–3 cm posterior to the tongue tip sensor), tongue dorsum (approximately 2–3 cm posterior to the tongue blade sensor) and soft palate. Each of the sensors provides x and y coordinates. Data recorded from each sensor are sampled at 500 Hz. Figure 1 shows the location of EMA sensors.

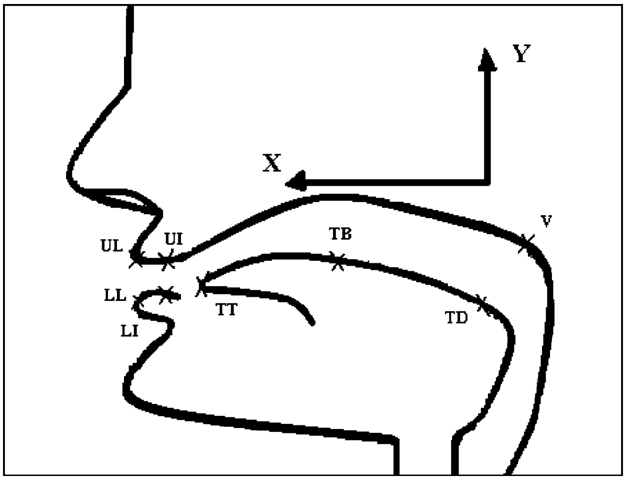


Fig. 1 Position of EMA sensors in x and y coordinates

In our experiments, we use the x–y coordinates of the upper and lower lip (UL, LL), lower incisor (LI), tongue tip (TT), tongue blade (TB), tongue dorsum (TD), and velum (V).

To obtain an acoustic representation, we apply the logarithm of the energy in the Hanning critical band filter banks based on bark scale (LHCB). The bandwidth of any filter in the filterbank is one bark [16]. Our experimental database includes 460 sentences from a single female speaker of British English (subject ID “fsew”, southern dialect) and a male speaker of British English (subject ID ‘‘maps’’, northern dialect) from the MOCHA database. In all experiments, we use 70% of the first data for training and the rest for testing.

Acoustic signal representations used in our experiments are LHCB. An LHCB representation vector containing 18 parameters is extracted from one speech frame. Frames length is 320 samples with 160 overlapping samples. LHCB features lie in the range between [0, 1].

The EMA data streams were down-sampled to 100 Hz to synchronize with parameters of LHCB. The range for each dimension of EMA is normalized to [0, 1].

The algorithm used to calculate LHCB is similar to the algorithm used to calculate Mel Frequency Cepstral

Coefficients (MFCCs). The fundamental difference between MFCC and LHCB is the nonlinear scale that is chosen for the frequency distribution of filters. LHCB and MFCC use the bark scale and mel scale, respectively [16]. The bandwidth of each filter in LHCB filterbank is one bark. Pervious works show that in neural network recognition systems, LHCB features can work more efficient than those of MFCC [12, 16–18]. Therefore, we use LHCB.

LHCB is calculated as follows:

1. Segment into frames of length $N = 320$ samples and remove the DC component.
2. Calculate short time Fourier Transform (STFT) of each frame $X(k)$ and calculate the spectral power $|X(k)|$.
3. Filter the spectral power using the square Hanning filterbank. For $0 \le k \le M$, the DFT of a Hanning filter is $\psi(k)$.

$$\psi(k) = \begin{cases} 0.5 - 0.5\cos\left(\frac{2\pi k}{M}\right), & 0 \le k \le M \\ 0, & \text{otherwise} \end{cases}, \tag{1}$$

where M is total number of filters ($M = 18$).

4. Calculate the logarithm of the output energy E_j of each filter.

$$E_j = \sum_{k=1}^{N} \left|\psi_j(k)\right|^2 |X(k)|^2, \quad j = 0, 1, \ldots, 18 \tag{2}$$

and

$$C_j = \log\left(1 + E_j\right) \tag{3}$$

5. LHCB features are obtained as follows:

$$c_m = \sum_{j=1}^{M} \ln(1 + E_j) \cdot \cos\left(m\left(\frac{(2j-1)}{2}\right)\frac{\pi}{M}\right); \quad 1 \le m \le L \tag{4}$$

where L is number of coefficient in the cepstrum domain ($L = 15$).

2.2 Motor-articulatory feedback theory

Motor-articulatory feedback theory is a neural-based theory that explains the reason of Alphabetic system disorder in phonological dyslexia. Dyslexia is a disability characterized by difficulty with reading text. This disorder includes at least two prominent subtypes: surface dyslexia (individuals cannot correctly utter the irregular words) and phonological dyslexia (individuals cannot correctly utter nonwords) [19, 20]. The latter is more common [19].

Phonological dyslexia is diagnosed in individuals who cannot use the Alphabetic system (learning the speech sounds that are related with letters), so they cannot correctly utter nonwords.

Indeed, patients have a problem in making a connection between sounds and Alphabetic symbols [13]. This reading disorder might be related to different neuropsychological or neurobiological pathologies [20]. Many different theories have attempted to explain the disorder, one of which is motor-articulatory feedback theory.

According to the theory, awareness of the positions and movements of articulatory system (lip, tongue, and jaw) would allow normal individuals to parse a word into its component phonemes. In phonological dyslexia, patients are not aware of the positions and movements of the articulators and are unable to utter a specific word [21]. This implies that there is feedback between the articulatory system and the brain in normal individuals. By using this feedback, better speech perception is probable and individuals could utter a word correctly. In other words, normal individuals have a bidirectional association between heard acoustic signals and articulatory movements. This association is controlled by brain. Inspired by this human perception operation, we propose a bidirectional neural network model for acoustic-articulatory inversion mapping.

2.3 Attractor dynamics

Suppose that an m dimensional discrete signal $\bar{s}(p)$ is trained to an auto-associative neural network, the structure of which is shown in Fig. 2. Suppose that the order of samples is not concerned, that is, $\bar{s}(p)$, $\bar{s}$ are considered as P samples in the input space which are trained to the network. At first, the activation function of neurons is supposed to be hard limit step function. In such a network, each neuron forms a hyperplane in its input space (the output space of the previous layer), and the input space is quantized by these hyperplanes. Every area constructed by quantizing hyperplanes is indicated with a unique binary code, i.e. if a sample in an area is affected by noise but it is still in the same area, the changes will not come into sight

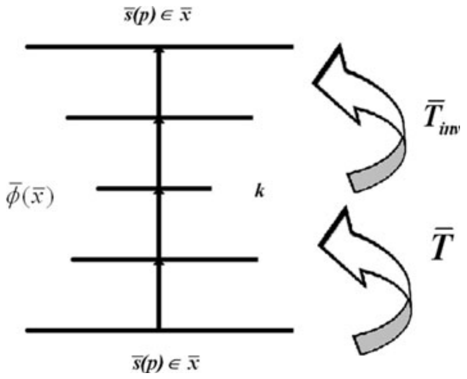


Fig. 2 Structure of an auto-associative neural network used to extract principal components

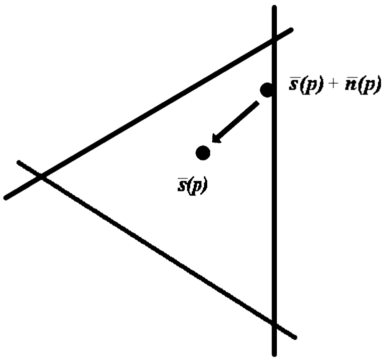


Fig. 3 The noisy sample, which is still in the region, seems to the next layer the same as the clean

in the next layer (Fig. 3). This quality is gained by the generalization of one point to an area (interpolation).

Practically, we decompose the input signal into its components using a series of basis functions $\varphi_j(\bar{x})$. Here, $\varphi_j(\bar{x})$ is a hyperplane corresponding to *j*th neuron which would be fired (1) or not fired (0).

In this case, some important issues must be considered: first, the position of each hyperplane should be determined so that the input sample would be placed in the center of the area to be able to tolerate the maximum disturbance. Second, the hyperplanes should be placed in a way that they can discriminate all samples in the input space, and this discrimination in the last layer leads to reconstructing the input samples with the least error.

Now, the activation functions of neurons are supposed to be a soft nonlinear function such as sigmoid, for instance. Therefore, the back-propagation algorithm can be used to train the network, and the boundaries will be soft and fuzzy. The output of each neuron is a continuous value dependent on the position of the $\bar{s}(p)$ in the input space and each input sample $\bar{s}(p)$ has an indication in each layer with these neurons.

At first, we suppose that only one sample is trained to the network with soft activation function. Only one neuron in the hidden layer is enough to indicate this sample, and for more neurons, their output will become equal after training [18].

In this case, the soft (fuzzy) hyperplanes are set in a way that the network achieves better representation for the sample, and the sample is reconstructed in the output space with nearly zero error. Besides, a cluster will be made around the attractor, which is constructed by soft hyperplanes. The input space is interpolated with these hyperplanes and every other point $s'(p)=\bar{s}(p)+\bar{n}(p)$ ($\bar{n}(p)$ indicates a noisy pattern) in this space will be projected on this unitary component. However, the value of hidden layer neuron for this noisy pattern is less than that for the original one. Due to the interpolation realized by the unitary kernel function $\varphi(\bar{x})$, this value is dependent on the distance and similarity between the noisy sample and the original one.

$$\bar{s}(p) \Rightarrow \varphi_{\max}(\bar{x}) \tag{5}$$

$$\bar{s}(p)+\bar{n}(p) \Rightarrow \varphi(\bar{x}) < \varphi_{\max}(\bar{x}) \Rightarrow \tilde{s}(p)+\tilde{n}(p) \tag{6}$$

where $\tilde{s}(p)+\tilde{n}(p)$ (clean signal + noisy pattern) is the final output of hidden layer. Corresponding to the original pattern here named $\bar{s}(p)$, the kernel function $\varphi(\bar{x})$will be maximum; $\varphi_{\max}(\bar{x})$. Now, if the sample is added to noise, $\bar{s}(p)+\bar{n}(p)$, the value of $\varphi(\bar{x})$ will be less than $\varphi_{\max}(\bar{x})$. This value of hidden layer results in an output which can be considered as $\tilde{s}(p)+\tilde{n}(p)$ where $\|\tilde{n}(p)\| < \|\bar{n}(p)\|$, that is, *the noise is reduced. The noise reduction happens because there is no unit in the network to represent the nonlinear principal components of the noise, but such a unit exists for the pattern. Therefore, the noise will be filtered nonlinearly. This output is again given to the input, and by cycling the noise will be reduced more and more. In every cycle, the noisy pattern moves toward the original pattern for one step. The cycling will be continued until the noise value is less than a threshold. The experimental results confirm the applicability of these facts* [17].

We consider a test. The test was applied to show how a bidirectional connection makes attractors. The model consists of two standard reverse-network structures. Each of forward and reverse parts has two hidden layers. First layer contains 32 neurons, second layer contains 64, and output layer contains three neurons. The structure of the model is shown in Fig. 4.

In the forward part, the output layer trains the code for each sample. The reverse part reconstructs the samples. The attractors behavior is considered. Figure 5 shows the nonnoisy inputs and outputs. The results show that the noisy inputs recognition performance is acceptable even for samples that are nonrecognizable by human (Fig. 6).

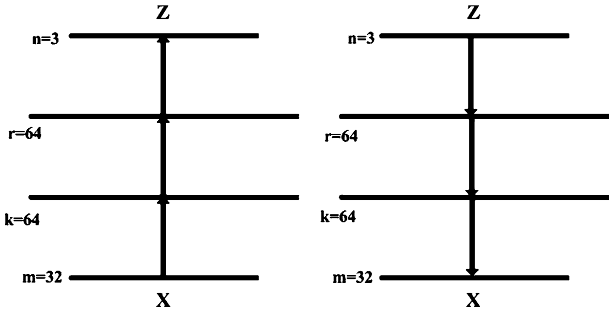


Fig. 4 The structure of an auto-associative BNN

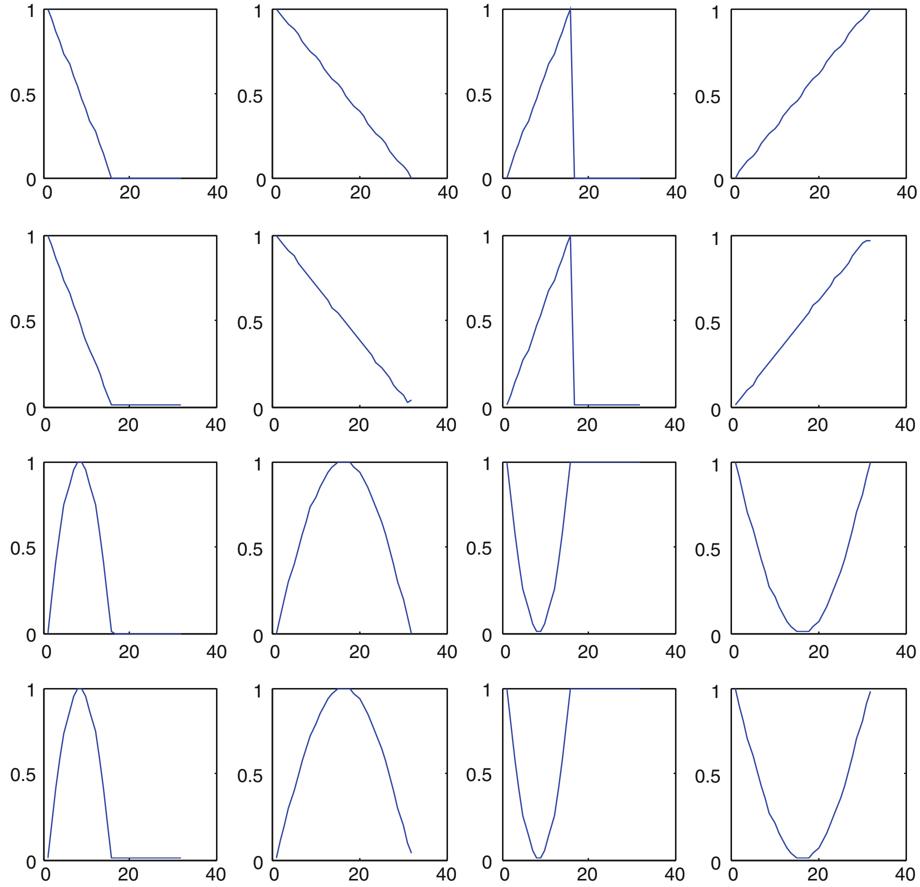


Fig. 5 Network inputs and outputs. *Rows 1 and 3* show the inputs without noise. *Rows 2 and 4* show the outputs

2.4 Feed forward network model

The first neural network model is a standard feed-forward network (FFN) that is designed in comparison with the bidirectional model. The FFN consists of one hidden layer. After 40 iterations, best mean square error (MSE) of the output in the model is obtained, while the hidden layer contains 92 neurons. The model is trained with the resilient backpropagation algorithm. Input vectors of the model are LHCB representations, and output vectors are 14 features from 14 channels of EMA [22].

2.5 Proposed bidirectional neural network model

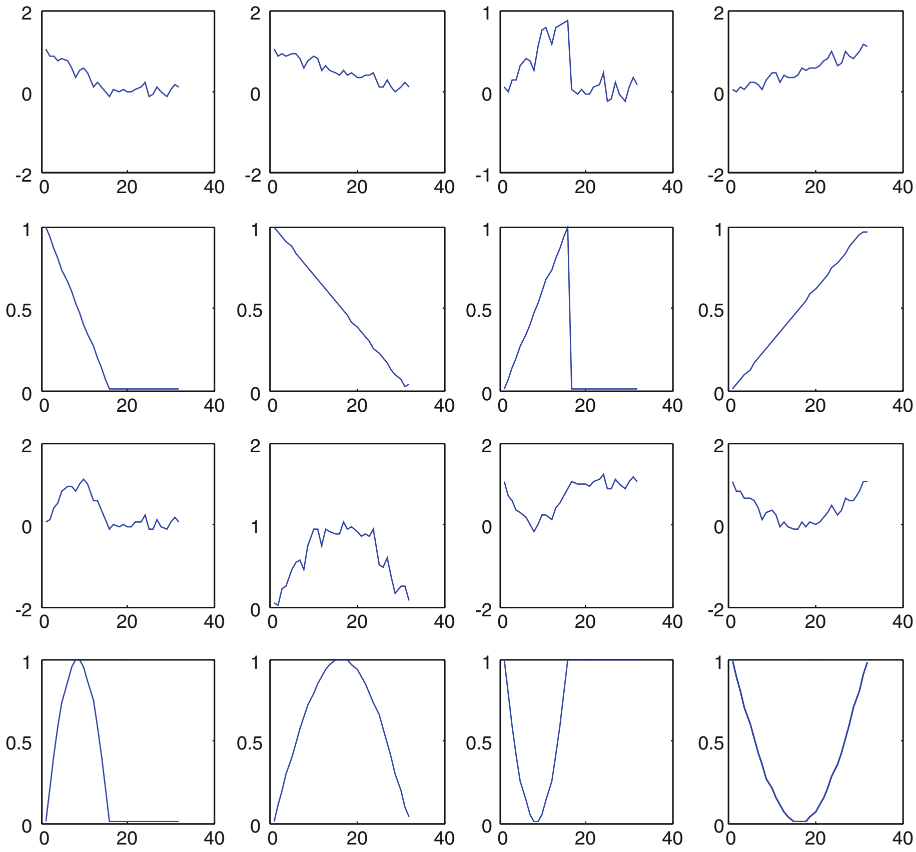
The second model that is proposed is a bidirectional (forward–reverse) model. The model consists of two standard reverse-network structures and attempts to mimic the performance, flexibility, correctness, and reliability of the human auditory system. Each parts of the model is trained independently. In the forward stage, we have a nonlinear mapping from LHCB representations to EMA features. Using the reverse part, we provide a nonlinear mapping from EMA features to LHCB representations. A general structure of the two networks is shown in Fig. 7. As mentioned in Sect. 2.4, the forward part uses one hidden layer perceptron with 92 neurons. The reverse network is designed with two hidden layers to map 14 EMA features to 18 LHCB representations. To obtain optimized neurons in the first layer, we examine between 1 and 128 neurons. In other words, we set the second layer 64 neurons and vary the number of neurons in the first layer. With the best neuron selection for the first layer, we can examine second layer. In our model, the first layer contains 95 neurons, and second layer contains 92 neurons. Just like the FNN, the resilient backpropagation algorithm is used for optimizing the error function [22].

The outputs of forward section are passed as input to the reverse network inputs and the outputs for the reverse network are passed back to the forward section. After performing six rotations between the forward and reverse parts, the EMA and LHCB parameters are fixed.

2.6 Recognition model

The TDNN is used for phoneme recognition and is based on the articulatory and acoustic parameters obtained in Sects. 2.4 and 2.5. A TDNN is a dynamic artificial neural network whose input, output, or both include not only current data values but also past and future values.

Fig. 6 Network inputs with 0 dB noise. *Rows 1 and 3* show the inputs. *Rows 2 and 4* show the outputs



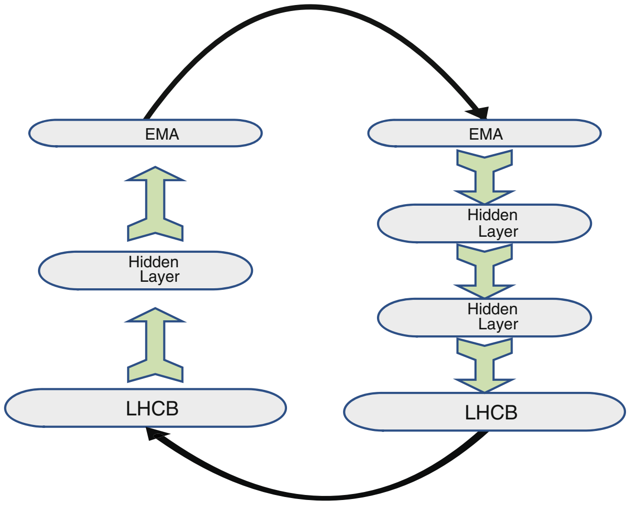


Fig. 7 General structure of BNN model training and testing process for the acoustic-articulatory inversion mapping

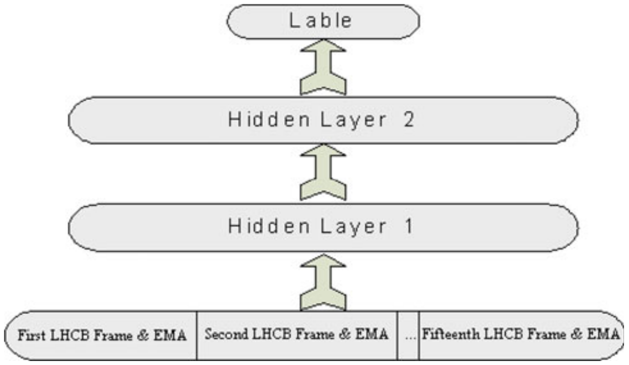


Fig. 8 A general structure of the TDNN

In our proposed TDNN, only the inputs contain information about the past and future. The model includes two hidden layers. Like the neuron selection method in the Sect. 2.4, best model is approximated. The first layer contains 93 neurons, and the second layer contains 70 neurons. A resilient learning algorithm for the optimization of the backpropagation error function is used to train the model [23]. Figure 8 shows a general structure of a TDNN model. The model uses 15 elements of LHCB representations and EMA values as input. In other words, the input includes one present element and fourteen past and future vectors, which contain both LHCB and EMA channel values. The TDNN recognition model learns various words in terms of their energy functions and articulatory vectors. The addition of temporal LHCB and EMA information allows the TDNN recognition model to learn words by sentence context. The FFN and BNN models prepare different mappings of EMA channels as auxiliary input in the phone recognition model.

3 Results

The results in this section are obtained by a model, which is trained and tested using the female speaker's data. Here, we use 70% of the first EMA data for training and the rest for testing. For comparison, a base TDNN that only uses LHCB as input is introduced. Its accuracy on the testing data is 53.11%. The TDNN that uses LHCB and EMA as input is applied, and accuracy on the testing data is improved up to 68.73%. Our first experiment uses the FFN to produce EMA from LHCB. The network uses the EMA output as auxiliary data in addition to the acoustic representation for phone recognition. The accuracy of this system is 62.09%. The proposed BNN model is used for inversion mapping. When the EMA output of this model is passed as auxiliary data to the TDNN model, the accuracy of phone recognition rate is improved up to 64.18%. Passing both the EMA output data and LHCB input data of BNN to the phone recognition model raises the recognition accuracy up to 64.91%. Table 1 shows the phone recognition accuracy in different models. As can be seen in Table 1, the accuracy is increased about 3% from 62.09 to 64.91%. Here, we use "McNemar's Test" to determine whether the recognition results can consider to be significantly different from one another. The McNemar significance level reflects the probability of the hypothesis that the differences between two classification results occur by chance. We set the threshold of the significance level to be 0.05, which means that the differences are considered as statistically significant if the probability of the differences occurring due to chance is less than 0.05 [24]. Here, the significance level of the McNemar's test is less than 0.001. Therefore, it indicates significant difference between these results.

Table 1 Comparison of recognition results with different models

Recognition model	Recognition accuracy
TDNN train with LHCB of MOCHA (fsew) and test with LHCB of MOCHA (fsew)	53.11
TDNN train with LHCB and EMA of MOCHA (fsew) and test with LHCB and EMA of MOCHA (fsew)	68.73
TDNN train with LHCB and EMA of MOCHA (fsew) and test with LHCB of MOCHA (fsew) and the auxiliary data(EMA) that is obtained from FFN model	62.09
TDNN train with LHCB and EMA from MOCHA (fsew) and test with LHCB of MOCHA (fsew) and the auxiliary data(EMA) that is obtained from BNN model after six rotations	64.18
TDNN train with LHCB and EMA of MOCHA (fsew) and test with auxiliary data (LHCB and EMA) that is obtained from BNN model after six rotations	64.91

3.1 Using other speaker features (male, northern accent)

In this section, we use 70% of the first EMA and corresponding acoustic data from the female speaker for training process. Besides, we use 30% of the last 460 sentences from the single male speaker for testing process. This database passes to the network that is trained by the female speaker features. In fact, male speaker features are passed to the inverse mapping network that is trained with the female features. After six rotations, the features pass to the phone recognition network. The principles of this method could be used in speaker-independent task if suitable databases were available. The results are in Table 2. In the male speaker, when articulatory feature is used beside the acoustic representation, the changes are minor. This is possibly because of the difference in the sex of speakers. It means that when they say a same sentence, their articulator's positions are different.

In the Table 2, we train the model with speech representation of the woman data (fsew). Then, we test the model with speech representations of the man (maps). Results show that the accuracy is 46.4%. In the second row, we use speech and articulatory representations of the woman to train the recognition model. Then, we use speech

Table 2 Comparison of recognition results with different models

Recognition model	Recognition accuracy
TDNN train with LHCB of MOCHA (fsew) and test with LHCB of MOCHA (maps)	46.4
TDNN train with LHCB and EMA of MOCHA (fsew) and test with LHCB and EMA of MOCHA (30% of the last "maps")	47.06
TDNN train with LHCB and EMA of MOCHA (fsew) and test with LHCB of MOCHA (30% of the last "maps") and the auxiliary data(EMA) that is obtained from FFN model when inputs are (maps)	47.32
TDNN train with LHCB and EMA from MOCHA(fsew) and test with LHCB of MOCHA (30% of the last "maps") and the auxiliary data (EMA) that is obtained from BNN model after six rotations when inputs are (maps)	48.8
TDNN train with LHCB and EMA of MOCHA(fsew) and test with auxiliary data (LHCB and EMA) that is obtained from BNN model after six rotations when inputs are LHCB (maps)	49.11

and articulatory representations of the man to test. The accuracy is raised up to 47.06%. In the third row, the training process is the same as the second row. However, in testing process, we use the articulatory features that are obtained from FFN mode, and the accuracy is improved again. In other words, although we eliminate the man data in the testing process, we see an improving in the results and the accuracy is raised to 47.32%. Again, we use the output of BNN model in acoustic-articulatory inversion mapping. The accuracy is again improved and raised to 48.8%. The point of this accuracy improvement is probably attractors. It means that the BNN model sends the Man's speech representations to the Woman's speech representations attractors. Finally, in the last row, the speech and articulatory representations are used in the BNN model at the same time. The accuracy is 49.11%. After eliminating man's speech and articulatory data in the process, we witness new growth. It shows that BNN models probably filter speech and auditory data. In the speaker-independent recognition, using such BNN model might help us to eliminate other speakers' data in testing process as well as getting better speech recognition.

4 Conclusion

The human perception has a bidirectional structure. When an individual utters a word, feedback from the acoustic signal is emitted from the speech organs. The articulatory movements of these organs are induced by feedback between heard acoustic signals and the sensing of articulatory movements. This learning process is placed in the best training mode (attractors). The goal of this paper was to establish this theory in the structure of ANNs and introduce a new BNN. The feedback between articulators and heard acoustic signals is used in the structure of the bidirectional model as forward and reverse networks. The method is compared with an FFN model. When used for speech recognition, the proposed model obtains higher accuracy using a BNN model in the acoustic-articulatory inversion process than it does using a standard FFN and TDNN. In other words, We want to investigate the ability of articulatory information to improve speech recognition. Here, we use a basic recognition system and try to improve the recognition rates by using articulatory and acoustic features jointly. Our research shows that using articulatory features along with acoustic representations in neural network systems could lead to better results. It happens even when we use acoustic and EMA data of male speaker as test data in a system that is trained by female speaker. Besides, we consider an acoustic-articulatory inversion mapping system. We use this system to map the acoustic and articulatory data of male speaker to those of female speaker. It shows that probably a person's knowledge of their own acoustic and articulatory information can help them to learn corresponding speech that is produced by other people. This method can be used for speaker-independent recognition if the suitable databases were available. In the future, method can be applied to inversion mapping methods, such as HMM-based methods and others.

Acknowledgments We would like to thank Alan wrench, Mark Hasegawa-Johnson, and Sarah Borys for the useful discussions and other help along the way.

References

1. Wrench A, Richmond K (2000) Continuous speech recognition using articulatory data. In: Proceedings of the ICSLP, Beijing, China, pp 145–148
2. Frankel J, Richmond K, Simon K, Taylor P (2000) An automatic speech recognition system using neural networks and linear dynamic models to recover and model articulatory tracks. In: Proceedings of ICSLP, vol 4, pp 254–257
3. Deng L, Erler K (1992) Structural design of hidden Markov model speech recognizer using multivalued phonetic features: comparison with segmental speech units. J Acoust Soc Am 92(6):3058–3067
4. Richmond K (2006) A trajectory mixture density network for the acoustic-articulatory inversion mapping. In: Proceedings of interspeech, Pittsburgh, USA
5. Richmond K (2007) Trajectory mixture density networks with multiple mixtures for acoustic-articulatory inversion. In: Proceedings of NOLISP, Paris, France, pp 263–272
6. Richmond K (2007) Multitask learning perspective on acoustic-articulatory inversion. In: Proceedings of interspeech, Antwerp, Belgium, pp 2465–2468
7. Richmond K, King S, Taylor P (2003) Modeling the uncertainty in recovering articulation from acoustics. Comput Speech Lang 17:153–172
8. Toda T, Black A, Tokuda K (2004) Acoustic-to-articulatory inversion mapping with Gaussian mixture model. In: Proceedings of 8th international conference on spoken language processing, Jeju, Korea, pp 1129–1132
9. Hogden J, Lofqvist A, Gracco V, Zlokarnik I, Rubin P, Saltzman E (1996) Accurate recovery of articulator positions from acoustics: new conclusions based on human data. J Acoust Soc Am 100(3):1819–1834
10. Zhang L, Renals S (2008) Acoustic-articulatory modeling with the trajectory HMM. IEEE Signal Process Lett 15:245–248
11. Kello CT, Plaut DC (2004) A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters. J Acoust Soc Am 116(4):2354–2364
12. Yazdchi MR, Seyyedsalehi SA, Zafarani R (2007) A new bidirectional neural network for lexical modeling and speech recognition improvement. Scientica Iranica 6:571–578
13. Heilman KM, Voeller K, Alexander AW (1996) Dyslexia: a motor-articulatory feedback hypothesis. Ann Neurol 39(3): 407–412
14. Waibel A, Hanazawa T, Hinton G, Shikano K, Lang KJ (1989) Phoneme recognition using time-delay neural networks. IEEE Trans Acoust Speech Signal Process 37(3):324–329
15. Wrench A (2000) A multi-channel/multi-speaker articulatory database for continuous speech recognition research. Phonus 5:1–13

16. Nejadgholi I, Seyyedsalehi SA (2007) Nonlinear normalization of input patterns to speaker variability in speech recognition neural networks. Neural Comput Appl 18(1):45–55
17. Dehyadegary L (2005) Noisy and distorted speech enhancement using neural networks. M.S. Thesis, Department of Biomedical Engineering, Amirkabir University of Technology (in Persian)
18. Seyyedsalehi SA, Nejatgholi I, Tohidkhah F (2004) Feed forward neural networks recognition performance improvement using bidirectional processing. In the research project final report, Biomedical Engineering Faculty, Amirkabir University of Technology (in Persian)
19. Castles A, Coltheart M (1993) Varieties of developmental dyslexia. Cognition 47(2):149–180
20. Rapcsak SZ, Beeson PM, Henry ML, Leyden A, Kim E, Rising K, Andersen S, Cho H (2008) Phonological dyslexia and dysgraphia cognitive mechanisms and neural substrates. Cortex 45(5):575–591
21. Behbood H, Fallahnezhad M, Seyedsalehi SA, Gharibzadeh S (2010) Improving phonological dyslexia using electrical stimulation in articulatory system. J Neuropsychiatr Clin Neurosci 22(3):352
22. Behbood H, Seyyedsalehi SA, Tohidypour HR (2010) A new bidirectional neural network model for the acoustic-articulatory inversion mapping. Speech Prosody 2010, Chicago, USA
23. Behbood H, Seyyedsalehi SA, Tohidypour HR (2010) A novel feature extraction for neural—based modes in acoustic-articulatory inversion mapping. Speech Prosody 2010, Chicago, USA
24. Gillick L, Cox S, (1989) Some statistical issues in the comparison of speech recognition algorithms. ICASSP 1989, vol 1. Glasgow, UK, pp 532–535