A New Bidirectional Neural Network Model for the Acoustic-Articulatory Inversion Mapping For Speech Recognition

Hossein Behbood, Seyyed Ali Seyyedsalehi, Hamid Reza Tohidypour*

Amirkabir University of Technology, Dep. of Biomedical Engineering, Tehran, Iran
***Hossein_1779@aut.ac.ir**

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-inverse 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 the articulatory features. In comparison with a standard model, the output of bidirectional model as auxiliary data in phone recognition process, increases the accuracy up to approximately 3%.

Index Terms: 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, this data generates excellent results for an enhanced speech recognition, analysis, and synthesis [1]. For example, some efforts for using articulatory features in quest of having better speech recognition can be observed in [1,2].

Before the first attempts to produce the suitable articulatory gesture with reliable and precise equipment, researchers used vocal-tract models or linguistic rules [3]. Today we have continuous smooth data measuring with sophisticated equipment. However, the use of such equipment is impossible due to their high cost and complications. Therefore, the efforts are focused on a method to estimate the articulatory features from the acoustic signals. 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], TMDNs with multiple mixtures model [5], multi task 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 system to recovery articulatory movements from speech acoustics [10]. Recently, more attempts are reported about using Artificial Neural Networks (ANNs) [11, 5]. Moreover, human perception system shows that it has a bidirectional structure [19-24]. Therefore, in this paper we focused 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, primary recognition is accomplished by reliable regions. Unreliable regions are corrected afterwards. This action is iterated until the recognition is completed and the primary recognition is modified to the final recognition (fix point).
One of the prominent theories that explains these connections is Motor Articulatory Feedback theory [12]. 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 having 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 phone recognition process [13]. In this study, we successfully employ a special structure of 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 to Electromagnetic articulography (EMA) features and then pass the outputs of these models 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. Motor-Articulatory Feedback theory
Motor-Articulatory Feedback theory is a neural-based theory that explains 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 can't correctly utter the irregular words) and

phonological dyslexia (individuals can't correctly utter non-words) [14, 15] which the latter is more common [14].
Phonological dyslexia is diagnosed in individuals who can't use the Alphabetic system (learning the speech sounds that are related with letters), so they can't correctly utter non-words.
Indeed, patients have problem in making connection between sounds and Alphabetic symbols [12]. This reading disorder might be related to different neuropsychological or neurobiological pathologies [15]. 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 and are unable to utter a specific word [16]. Therefore, there would be a feedback between articulatory system and brain of normal individuals and by using this feedback, better speech perception is probable and they 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 of this human perception operation, we propose a bidirectional neural network model for the acoustic-articulatory inversion mapping.

3. Speech database and pre -processing

The Multichannel Articulatory (MOCHA) database consists of corpus of 460 TIMIT sentences of 40 different speakers [17]. This database includes acoustic signals, Laryngograph (LAR), Electropalatograph (EPG) and Electromagnetic articulography. Acoustic signals are recorded by sampling frequency of 16000 Hz samples per second. EMA sensors are connected to upper and lower lips, lower incisor (jaw), tongue tip (5-10mm from the tip), tongue blade (approximately 2-3cm posterior to the tongue tip sensor), tongue dorsum (approximately 2-3cm posterior to the tongue blade sensor) and soft palate. Each of the sensors provides x and y positions recorded from each sensor that samples at 500 Hz. Figure1 shows the location of EMA sensors.
In our experiments we use upper and lower lip, lower incisor, tongue tip, tongue blade, tongue dorsum and velum data in x and y coordinates. For acoustic representation, apply logarithm of energies in the Hanning type critical band filter banks (LHCB) representation [18]. Our experimental database includes corpus of 460 from one female speaker of British English (subject ID ''fsew'', southern dialect) in the MOCHA database. We use 70% of the 460 sentences and their parameters of EMA as training process and 30% for testing.
Acoustic signal representations used in our experiments are LHCB. Any representation vector containing 18 parameters is extracted from one speech frame which is logarithm of energies in the Hanning type critical band filter banks while bandwidth of any filter is one bark. 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].

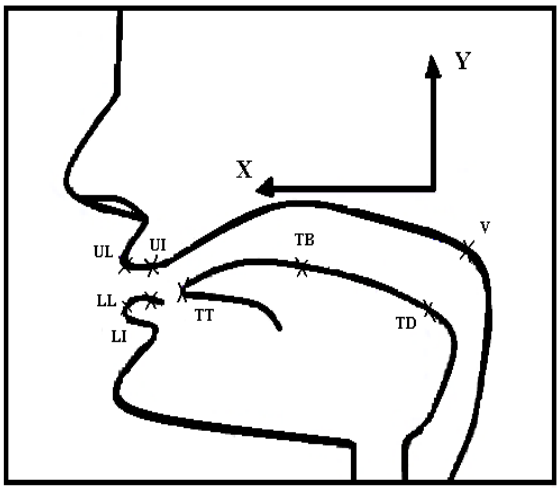

Figure1: Position of EMA sensors in x and y coordinates

4. Inversion models

For using Motor-Articulatory Feedback theory in conversion between acoustic representations (LHCB) and EMA features, we proposed a BNN model inspired from feedback between articulatory system and the acoustic energy of speech signals that are uttered by human. This feedback is controlled by brain as mentioned in section 2. Proposed BNN model consists of two parts: Forward part and reverse part. The former is used for a nonlinear mapping of LHCB representations to EMA features and the later provides a nonlinear conversion of EMA features to LHCB representations. To fix EMA and LHCB parameters, 6 rotation between the inputs and outputs are performed. In all models, training iteration process is interrupted in the best mode.

4.1. Feed Forward Network model

The first neural network model is a standard FFN that is designed in comparison with bidirectional model. The FFN consists of one hidden layer. Best MSE of output in this model as shown in figure 2 is obtained while hidden layer involves 92 neurons. The model is trained with resilient backpropagation algorithm. Input vectors of the model are LHCB representations and output vectors are 14 features from 14 channels of EMA.

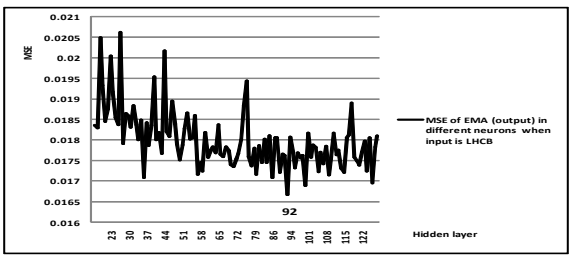

Figure2: MSE of EMA (output) in different amount of neurons in the hidden layer of FFN in 40 iteration

4.2. Bidirectional Neural Network model

Second model which is proposed, is a bidirectional (forward-reverse) model. The model is consist of two standard networks in reverse structures and tries to get closer to the performance, flexibility, correctness and reliability of the human auditory system. At forward part we have a nonlinear mapping from LHCB representations to EMA features. Afterward, by using a reverse part, we provide a nonlinear mapping from EMA features to LHCB representations. A general structure of the two networks is shown in Figure3. As mentioned in section 4.1, the forward part uses one hidden layer perceptron with 92 neurons. The reverse part is designed in two hidden layer because we want to map 14 EMA features to 18 LHCB representations and this mapping is harder than forward part. To obtain optimized neurons in the first layer, we examine from 1 to 128 neurons. In other words: Put second layer 64 neurons and test the different amount of neurons in the first layer (shown in part A in figure 4).Then with the best neuron selection for this layer, examine second layer (Shown in part B in figure 4). Therefore, first layer involves 93 neurons and second layer involves 70 neurons. Just like the FNN, resilient backpropagation algorithm is used for optimizing the error function.

The outputs of forward section are passed as reverse part inputs and vice versa, respectively. After performing 6 rotation between forward and reverse parts, the EMA and LHCB parameters is fixed.

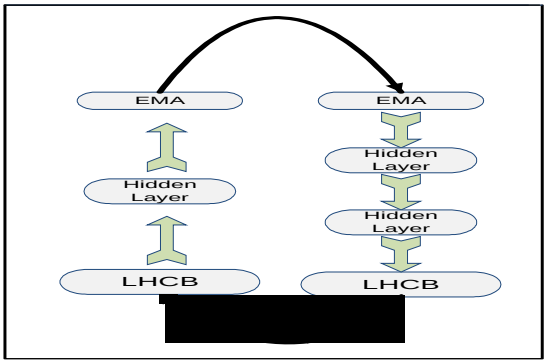

Figure 3: General structure of BNN model for the acoustic-articulatory inversion mapping

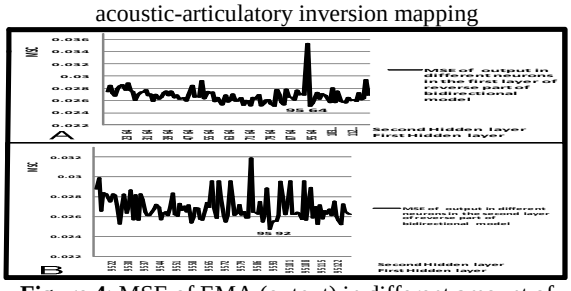

Figure 4: MSE of EMA (output) in different amount of neurons in the first and second hidden layers of reverse part of BNN model in 40 iteration

5. Recognition model

A TDNN is used for phone recognition based on the articulatory and acoustic parameters are obtained in section 4.1 & 4.2. A TDNN is a dynamic model of artificial neural networks which inputs or outputs vectors or both including not only the current values but also the past and future values.

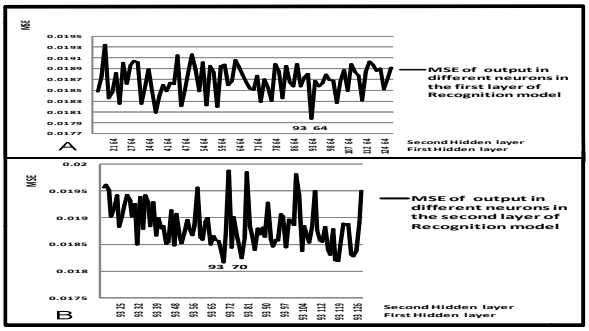

Figure5: MSE of output in different amount of neurons in the first and second layers of TDNN model in 40 iteration

In our proposed TDNN, there are just inputs swapping in time. The model includes two hidden layers. Like the neuron selection method on the section 4.2, best model is approximated. First layer involves 93 neurons (shown in part A of figure 5) and the second layer involves 70 neurons (shown in part B of figure 5). Resilient learning algorithm for optimization of error function in backpropagation structure is used. Figure 6 shows a general structure of TDNN model. The model uses fifteen elements of LHCB representations and EMA value. In fact, it uses not only each LHCB & EMA vector but also fourteen past & future vectors of LHCB and values of EMA channels. As a result of swapping in time over the LHCB & EMA frames, the TDNN recognition model learns various words in training process not only by their energy functions and articulatory vectors but also by sentence context which they are used in. Actually, FFN and BNN models prepare different mapping of EMA channels for applying as auxiliary input in the phone recognition model.

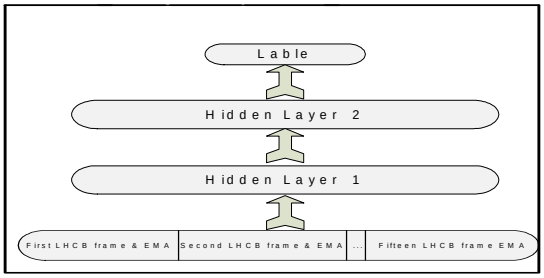

Figure 6: A general structure of Time Delay Neural Network

6. Experimental results

In comparing, a base TDNN is introduced that only use LHCB as input. The accuracy was 53.11%. When using the FFN model for producing EMA from LHCB and passing output (EMA) as auxiliary data beside the acoustic representation for phone recognition, the accuracy was 62.09%. By using proposed BNN model for this inversion mapping, and passing output (EMA) of this model as auxiliary data to TDNN model, accuracy of phone

recognition rate is improved up to 64.18%. Afterward, passing output (EMA) data and input data (LHCB) of BNN to the phone recognition model, the accuracy rate is increased again up to 64.91. Table 1 shows the comparison of phone recognition accuracy in different models.

Table 1. Comparison of recognition results with different models

Recognition model	Recognition accuracy
TDNN train with LHCB of MOCHA and test with LHCB of MOCHA	53.11
TDNN train with LHCB &EMA of MOCHA and test with LHCB &EMA of MOCHA	68.73
TDNN train with LHCB &EMA of MOCHA and test with LHCB of MOCHA and the auxiliary data(EMA) that is obtained from FFN model	62.09
TDNN train with LHCB &EMA from MOCHA and test with LHCB of MOCHA and the auxiliary data(EMA) that is obtained from bidirectional model after 6 rotations	64.18
TDNN train with LHCB &EMA of MOCHA and test with auxiliary data (LHCB & EMA) that is obtained from bidirectional model after 6 rotations	64.91

7. Conclusion and future work

Human perception system shows that it has a bidirectional structure. When a individual utters a word, it would be a feedback from acoustic signals emitted from speech organs and articulatory movements of these organs so with iteration in feedback between heard acoustic signals and the sense of articulatory movements, learning process is placed in the best training mode (fix point). The paper has aimed simply to establish this theory in the structure of ANNs and introduce a new BNN. Actually, the feedback between articulators and heard acoustic signals is used on the structure of bidirectional model as forward and reverse parts. The method is compared with a FFN model. It gets higher accuracy in speech recognition process when we use BNN model in acoustic-articulatory inversion process instead of standard FFN and pass the outputs as auxiliary data to the phone recognition network (TDNN). This method can be used for speaker-independent recognition if the suitable databases were available. Also this way could be used in any kind of inversion mapping methods such as HMM-based and others in the future .

9. Acknowledgments

We would like to thank Dr. Alan. Wrench for useful discussion and other help along the way.

References:

[1] Wrench, A., Richmond, K., "Continuous speech recognition using articulatory data", In: Proc, ICSLP, 2000

[2] Frankel, J ., Richmond, K., Simon, K ., Taylor, P. "An Automatic Speech Recognition System Using Neural Networks And Linear Dynamic Models To Recover And Model Articulatory Tracks", ICSLP, 2000

[3] Deng, L. & Erler, K., " 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, 1992

[4] Richmond, K., "A trajectory mixture density network for the acoustic-articulatory inversion mapping", In: Proc. Interspeech, Pittsburgh, USA, September 2006

[5] Richmond, K., " Trajectory Mixture Density Networks with Multiple Mixtures for Acoustic-articulatory Inversion", Springer, Volume 4885, 2007

[6] Richmond .k., "Multitask Learning Perspective on Acoustic-Articulatory Inversion", interspeech, 2007

[7] Richmond, K., King, S., Taylor, P.," Modeling the uncertainty in recovering articulation from acoustics", Computer Speech and Language, 153–172, 17, 2003

[8] Toda, T., Black, A., Tokuda, K., "Acoustic-to-articulatory inversion mapping with Gaussian mixture model", In: Proc. 8th International Conference on Spoken Language Processing, Jeju, Korea, 2004

[9] Hogden, J., Lofqvist, A., Gracco, V., Zlokarnik, I., Rubin, P., Saltzman, E., "Accurate recovery of articulator positions from acoustics: New conclusions based on human data", J. Acoust. Soc. Am. 100(3), 1819–1834, 1996

[10] Zhang, L., Renals, S., "Acoustic-Articulatory Modeling With the Trajectory HMM", Signal Processing Letters, IEEE, 2008

[11]. Christopher, T., Kelloa ., David C. Plaut.,"A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters", J. Acoust. Soc. Am. 116 ,2004

[12] Heilman, K. M., Voeller, K. et al,. "Dyslexia; A motor-articulatory feedback hypothesis", Annals of Neurology, 407- 412, 1996

[13] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., Lang, K.J.,"Phoneme recognition using time-delay neural networks", Acoustics, Speech and Signal Processing IEEE Transactions,1989

[14] Castles, A., Coltheart, M.," Varieties of developmental dyslexia. Cognition", 47(2), 149-80, 1993

[15] Rapcsak, S. Z, Beeson P. M, et al., "Phonological dyslexia and dysgraphia cognitive mechanisms and neural substrates". Cortex, 2008, Volume 45, Issue 5, 575-591

[16] Behbood, H., Fallahnezhad, M., Seyedsalehi, S. A., Gharibzadeh, S.," Improving Phonological Dyslexia using Electrical Stimulation in articulatory System", The Journal of Neuropsychiatry and Clinical Neurosciences, 2009

[17] Wrench, A.," A Multi-Channel/ Multi -Speaker Articulatory Database For Continuous Speech Recognition Research", Phonus, vol. 5, pp. 1-13,2000

[18] Nejadgholi, I., Seyyedsalehi, S.A.," Nonlinear normalization of input patterns to speaker variability in speech recognition neural networks", Neural Comput & Applic, Volume 18, Number 1 2007

[19] Yazdchi .M.R., Seyyedsalehi S. A., Zafarani. R. "A New Bidirectional Neural Network for Lexical Modeling and Speech Recognition Improvement ", Scientica Iranica, vol 6, pp571-578

[20] Korner,E.,Gewaltig,M.,Korner,U.,Richter,A.and Ro demann ,T. , "A model of computation in neocortical architecture", NeuralNetworks, 12(7-8),pp989-1005 (1999).

[21] Koerner, E., Tsujino, H. and Masutani, T. "A cortical-type modular neural network for hypothetical reasoning-I I.The role of cortico-cortical loop", Neural Networks, 10(5),pp791-814(1997).

[22] Korner,E, Matsumoto,G., "Corticalarchitecture self-referential control for brain-like computation", and Engineering in Medicine and Biology Magazine ,IEEE, Engine (5),pp121-133(2002). 21

[23] Ghosn, J., Bengio, Y., "Bias learning, knowledge. Neural Networks, IEEE Transactions on, sharing", (4),pp748-765(2003).

[24] Mesulam, M., From Sensation to Cognition, Brain, 121(6),pp1013-1052(1998).