A Novel Feature Extraction for Neural – based Modes in Acoustic-Articulatory Inversion Mapping

Hossein Behbood[*], *Seyed Ali SeyyedSalehi* , *Hamid Reza Tohidypour*
Amirkabir University of Technology, Department of Biomedical Engineering

[*]Hossein_1779@aut.ac.ir

Abstract:

Acoustic-articulatory inversion mapping is a process that converts the signal of acoustic data to articulatory features. Most research focused on finding the best model for this mapping process but less attention on finding appropriate representation of articulatory & acoustic signals. This paper suggests two feature extraction methods, including Logarithm of square Hanning Critical Bank filterbank & Discrete Wavelet Transform that have better operation in contrast with conventional feature extraction based on Mel-Frequency Cepstral coefficients. For inversion mapping process an standard feed forward neural network is used. Appling a Time Delay Neural Network for phone recognition. The results show the efficiency of two new feature extraction methods.

Index Terms: Discrete Wavelet Transform, Time Delay Neural Networks (TDNNs), MOCHA-TIMIT database, Acoustic- Articulatory Inversion Mapping, Logarithm of square Hanning Critical Bank filterbank (LHCB), Mel Frequency Cepstral Coefficients(MFCC)

1. Introduction

Combine articulatory features with acoustic signals as auxiliary data to improve speech recognition have been very common lately. First attempt focused on using vocal-tract models or linguistic rules data combined with acoustic features [3]. Afterward, the scientists product continuous smooth data measuring with precision and reliable equipments so the articualtory features that was gained from these equipments are applied. Fantastic results are reported in combination of articulatory features with acoustic representations for speech recognition, analysis and synthesis [1,2].

The use of these equipments data as usual is impossible. The high cost of the equipments and invasive method in the recording process are the obstacles for using these data in normal process. Therefore, the attempts are concentrated on estimate the articulatory movements from the acoustic signals. Estimating of articulatory movement from acoustic signals is called inversion mapping. Several estimation methods focused on obtained the best estimation models such as a Trajectory Mixture Density Networks (TMDNs) model [4], TMDNs with Multiple Mixtures model [5], using a multi task learning perspective [6], modeling the uncertainty in recovering articulation from acoustics [7], using Gaussian Mixture Model (GMM) [8], accurate recovery of articulator 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]. Nevertheless, there is less concentration on finding the best representation for inversion mapping process. The majority of works on this filed using Mel Frequency Cepstral Coefficient (MFCC) as representation feature because of the good efficacy and also easy extrication from acoustic signals with helping from HTK toolkit. But, In this article, we focused on finding better representation for inversion mapping process especially for using on ANNs. Log of square Hanning Critical Bank filterbank (LHCB) is one of feature extraction method that used in [18] for speech recognition. Wavelet Packet and reforming Discrete Wavelet Transform use for better extraction of representation features from speech signals is continue [20].

A standard Feed-forward Network (FNN) model is used for nonlinear inversion mapping. For comparison, we use a Time Delay Neural Networks (TDNNs) model. Ability of TDNN in speech recognitions has been proven [13]. Employing a special structure of TDNNs, use past & future inputs instead of using every input, individually. In all neural network structures, use Backpropagation Network (BPN) models with resilient optimization algorithm for minimizing error function. In all models, training iteration is interrupted in the best mode.

For each feature extraction methods a special recognition model is trained. In comparison between three kinds of feature extraction methods, the LHCB method gain best results. Discrete Wavelet Transform and MFCC are in the next rates. The results of this study show that must have fundamental reconsideration in using MFCC representations in ANN-based inversion mapping process.

2. 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 square Hanning Critical Bank filterbanks (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 that are extracted from one speech frame which are 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 parameters of LHCB. The range for each dimension of EMA is normalized to [0, 1].

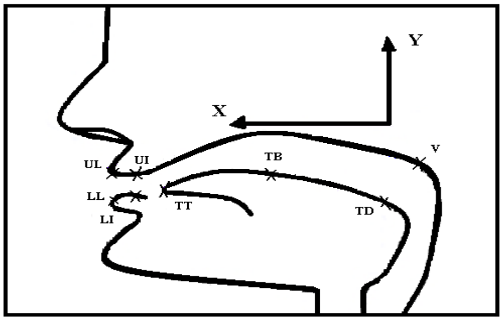

Figure 1: Position of EMA sensors in x and y coordinates

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

3. Feature Extraction Methods:

In this section MFCC and LHCB, which are based on Fourier transform, are presented. The other method is Discrete Wavelet Transform (DWT), which is based on multi resolution theories. Wavelet is used to analyze non-stationary signals [20].

3.1. Discrete Wavelet Transform:

After pre-processing, a frame of signal is chosen. Afterward, three DWT of speech signal are given. The DWT decomposition was performed up to level 6. We eliminate high frequency component of the first stage. That's beacause, after 8 KHz the information of speech signal is negligible. In this study, db4, db10 and db16 are used (each obtain 6 features). The advantage of choosing these mother wavelets is that increasing number of zero moments in wavelet, cause more oscillations in mother wavelets [21]. Therefore, DWT coefficients can better represent speech signals. By using Parsval theorem, the energy of each filters, is calculated, then we use them as train features.

3.2. Log of square Hanning Critical Bank filterbanks:

LHCB is abbreviation of Logarithm of square Hanning Critical Bank filterbanks that has the same distance according to bark scale. Algorithm of this method is as the same as MFCC, fundamental difference between MFCC and LHCB is non-linear scale, which has been chosen for the frequency and distribution of filters. As seen, for LHCB and MFCC parameters the Bark scale and Mel scale has been used respectively.

1-Choose a frame of acoustic signal with N=320 samples, & Remove dc amount of frame

2-Multiply frame by Hamming time domain windows

3-Calculate short time Fourier Transform of each frame, X(k) , Calculate spectral power $\left| X\ (k\ ) \right|$

4-Apply filter banks of square hanning to the spectral power. For $0 \le \mathrm{k} \le \mathrm{M}$, DFT of a hanning filter are $\psi(k)$. M is total number of filters (M=18).

5-Calculate the log energy output of each filter E_j for j=0,1,....,18 , then calculate logarithm of E_j

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

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

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

4. Models:

For mapping between the acoustic signal representations and EMA channels, a standard FNN model with one hidden layer is applied. This model is used for a nonlinear mapping of acoustic representation features to EMA channels. A TDNN model is used for phone recognition based on the articulatory and acoustic parameters is obtained in inversion mapping process. For each extraction method used specific inversion model and specific recognition model that obtained with find best model process. All models training iteration is interrupted in the best mode. Backpropagation Network (BPN) models with resilient optimization algorithm are used for minimizing error function.

4.1 Inversion Models:

For conversion between acoustic signal representations and EMA channels, we apply a neural - based model that consists of one hidden layer. For each representation extraction, obtain best amount of neurons for best MSE.

4.1.1. LHCB

The neural network model is a FNN. The FNN consist of one hidden layer. Best MSE of output in this model as shown in figure 5 obtain in 92 neurons. The model is trained with resilient backpropagation algorithm and training process is interrupted in the best mode. Input vectors of the model are LHCB representations and output vectors are 14 features from 14 channels of EMA.

4.1.2. Discrete Wavelet Transform

The inversion model consists of one hidden layer. Best MSE of output when input data is Wavelet as shown in figure 5 is obtained in 58 neurons. The model is trained with resilient back propagation algorithm and training process is interrupted in the best mode. Input vectors of the model are LHCB representations and output vectors are 14 channels of EMA.

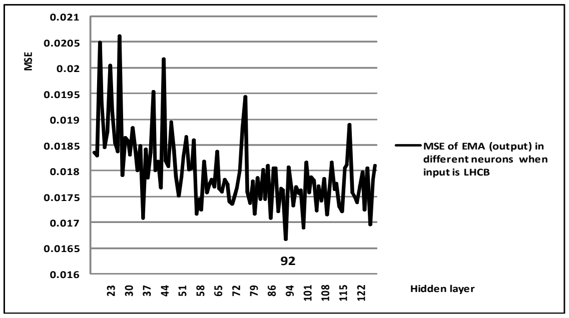

Figure 2: MSE of EMA (output) in different amount of neurons when input is LHCB

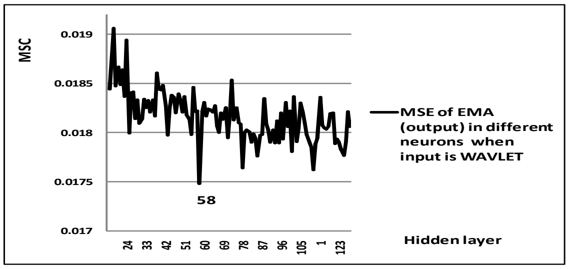

Figure 3: MSE of EMA (output) in different amount of neuron when input is Wavelet

4.1.3. MFCC

When MFCC representation used best MSE of output as shown in figure 4 obtained in 68 neurons. All models is trained with resilient backpropagation algorithm and training process is interrupted in the best mode.

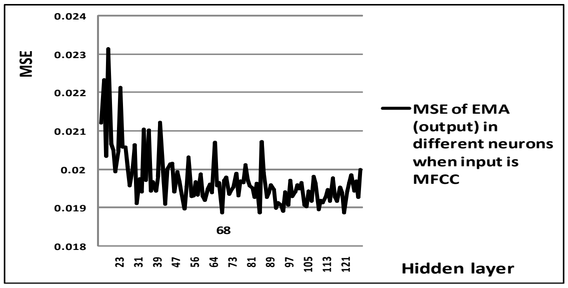

Figure 4: MSE of EMA (output) in different amount of neurons when input is MFCC

4.2. Recognition models:

A TDNN [19] used for phone recognition based on the articulatory and acoustic parameters is obtained in section 4. A TDNN is a dynamic model of artificial neural networks which inputs and outputs vectors or both including not only the current values but also the past & future values. In our proposed, resilient learning algorithm for optimization of error function in backpropagation structure is used. Figure 5 shows a scheme 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.

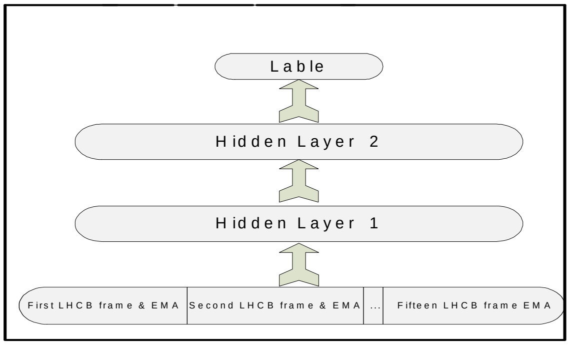

Figure 5: The Time Delay Neural Network for recognition

4.2.1 Optimized Recognition Models for LHCB features:

In this section, a optimized neural- based recognition model which used LHCB representation as input is introduced. The model includes two hidden layers. Like the neuron selection method on the section 4.2, best model for recognition is approximated. First layer involves 93 neurons (shown in part A of figure 6) and the second layer involves 70 neurons (shown in part B of figure 6).

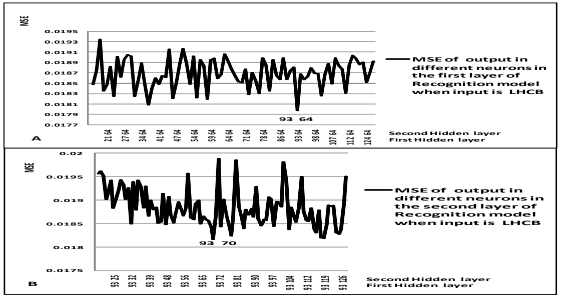

Figure 6: MSE of output in different amount of neurons in the first and second layers of recognition model in 40 iteration when input is LHCB

4.2.1 Optimized Recognition Models for DWT features

An optimized neural- based recognition model, which used DWT representations as input, is introduced. The model includes two hidden layers. Like the neuron selection method on the section 4.2, best model for recognition is approximated. First layer involves 78 neurons (shown in part A of figure 7) and the second layer involves 103 neurons (shown in part B of figure 7).

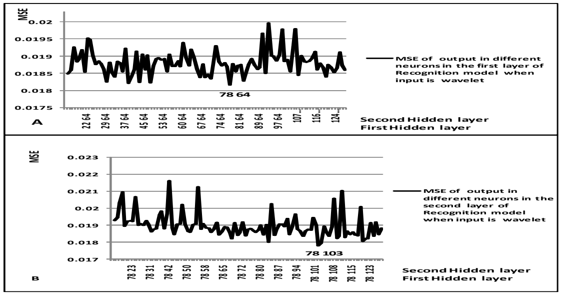

Figure 7: MSE of output in different amount of neurons in the first and second layers of recognition model in 40 iteration when input is wavelet

Table 1: Comparison of recognition results when used different representations as input in inversion model

Recognition modes	Recognition accuracy
TDNN train with LHCB &EMA of MOCHA and test with LHCB with axillaries data(EMA) that is obtained from related inversion model	62.09
TDNN train with Wavelet &EMA of MOCHA and test with Wavelet with axillaries data(EMA) that is obtained from related inversion model	61.38
TDNN train with MFCC &EMA of MOCHA and test with MFCC with axillaries data(EMA) that is obtained from related inversion model	59.72

4.2.1 Optimized Recognition Models for MFCC packet features

An optimized neural- based recognition model which used DWT representation as input is introduced. The model includes two hidden layers. Like the neuron selection method on the section 4.2, best model for recognition is approximated. First layer involves 93 neurons (shown in part A of figure 8) and the second layer involves 70 neurons (shown in part B of figure 8).

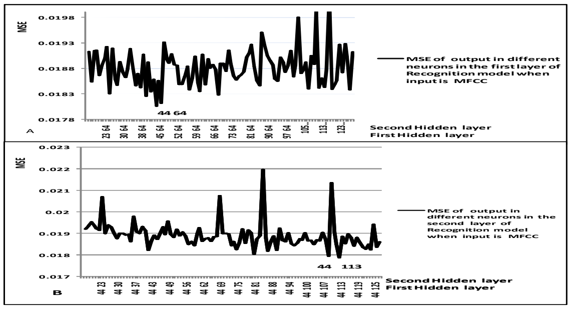

Figure 8: MSE of output in different amount of neurons in the first and second layers of recognition model in 40 iteration when input is MFCC

5. Experimental results:

When using the LHCB representations as input for inversion model and pass the output of this model (EMA) as auxiliary data to related phone recognition model, the accuracy was 62.09%. The DWT representations as input for inversion model and use the output of this model (EMA) as auxiliary data in related phone recognition model, the accuracy descend to 61.38%. Finally, use MFCC representations and use the output of this model (EMA) as auxiliary data in related phone recognition model. , the accuracy was 59.72%. Table 1 shows the comparison of phone recognition accuracy in different models.

6. Discussion:

This research shows the power of LHCB representations and DWT representations in comparison with common feature extraction (MFCC). This paper based on neural network models and there would be test in other inversion mapping models in the future.

7. 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

[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, Vo. 18, No. 1_ 2009

[19] 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

[20]Tufekci Z., Gowdy J.N., Gurbuz S., Patterson E.; 'Applied Mel-Frequency discrete wavelet coefficients and Parallel model compensation for noise-Robust Speech Recognition'; Speech Communication 48 (2006) 1294-1307

[21] Selesnick I. W. The double density DWT. In A. Petrosian and F. G. Meyer, editors, Wavelets in Signal and Image Analysis: From Theory to Practice. Kluwer, 2001.