International Journal of Soft Computing 11 (3): 140-144, 2016
ISSN: 1816-9503
© Medwell Journals, 2016

Multistage Enhancement of Multilingual Native Speaker Recognition in Neural Network on Mobile Data

Manju Patel and Rajeswari Mukesh
Hindustan University, Chennai, India

Abstract: Human is the most creative and destructive gift given by god to our nature. In order to show his creativity he invented a revolution able device mobile. As mobile phones have become ubiquitous and basic communications tools. It is most supporting tool for increasing the crime. Specially unauthorized multilingual voice calls or voice messages by nonnative person to break the border security. In order to recognize multilingual native speakers on mobile phones we have designed Multilingual Native Speaker Recognition System (MNSRS). To enhance MNSRS we use Multi Layer Feed Forward Neural Network (MLFFNN) in three stages. In the first stage we remove noise from the corrupted voice signal recorded by auto call recorder in mobile by using Hybrid Adaptive Neuro Fuzzy Inference System (ANFIS). We notice that ANFIS gives a faster convergence rate and improved SNR in poor acoustic environment and also increases MNSRS from 53.7-61.6% with 1950 MFCC values. In the second stage, we fused 1950 MFCC features with 13 LPC and 4 prosody features (pitch, intensity, third and fourth formants). With fused 1967 features recognition percentage of MNSRS increases from 61.6-65.5%. In the third stage, over fitting is the cause for degrading performance of MMNSRS.

Key words: Mobile corpus, ANFIS, MFCC, LPC, MNSRS, MLFFNN

INTRODUCTION

To show our feeling and emotions, we have a precious thing voice. It is the medium to convey the messages from one to others. In order to send voice massages electronically mobile phones have become ubiquitous and basic communications tools. Although cell phones have proven to be instrumental in reducing crime, they've also played a part in creating it. Specially unauthorized multilingual voice calls or voice messages from nonnative person to break the security. Mobile voice data has so many challenges like tower availability, signal strength, background noise and channel disturbance etc. In this study, multilingual corpus has designed on spice mobile by auto call recorder. Five multilingual native speakers (Hindi, Tamil, Malayalam, Kannada and Telugu) have taken for recording. Speakers are chosen who can speak all five languages as well English hence, the total spoken languages are 6 (Hindi, English, Tamil, Malayalam, Kannada and Telugu). All speakers speak text independent numbers (0-9) for training and (10-19) for testing.

In order to recognize multilingual native speakers on mobile phones we need a system to catch unauthorized and harmful voice call and massages. In Fig. 1, multi stages enhanced MNSRS is shown. In the first stage we have removed noise from the noisy voice signal by evolutionary hybrid technique ANFIS noise cancellation. As we know noise is an unwanted energy which interferes with the desired signal. It can be suppressed with adaptive filters using signal processing. But if the noise frequency is same as the original signal then sometimes it also eliminates the desired signal. Therefore, ANFIS noise cancellation is used which will not affect the desired signal. Neural networks recognize patterns and adapt themselves to cope with changing environments.

The basic principle of noise cancellation using neuro fuzzy is to filter out an interference component by identifying the nonlinear model between a measurable noise source and the corresponding immeasurable interference. The MATLAB command "ANFIS" (Adaptive Neuro Fuzzy Inference System) is used to demonstrate how noise cancellation can be applied as interference canceling in a signal (Suresh and Sundaravadivelu, 2008; Kelebekler and Inal, 2006). For ANFIS input audio signal contaminated with white and colored noise is taken from spice mobile by auto call recorder. The major advantage of ANFIS is it improves SNR with faster convergence rate and ease of implementation with higher accuracy (Hossain *et al*., 2013).

After ANFIS noise cancelation we extract auditory features 1950 Mel Frequency Cepstral Coefficient (MFCC) (Martinez *et al*., 2012). These features have taken as input to MLFFNN which classify five nativities of speakers (Hindi, Kannada, Malayalam, Tamil and Telugu) for

Corresponding Author: Manju Patel, Hindustan University, Chennai, India

MNSRS. MLFFNN network consists of multiple layers of computational units, usually interconnected in a feed-forward way. Each neuron in one layer has directed connections to the neurons of the subsequent layer. In this paper the units of these networks apply a sigmoid function as an activation function. Error back-propagation has used to calculate the error in each no of iteration then the output values are compared with the target to compute error by using gradient descent method (Fredrickson and Tarassenko, 1995; Praveen *et al*., 2013). The recognition percentage with 1950 MFCC of MNSRS increases from 53.7-61.6% with 20 hidden neurons.

In the second stage, we have extracted some more features 13 LPC and 4 prosody features (pitch, intensity, third and fourth formants). These features are fused with 1950 MFCC features (Hosseinzadeh and Krishnan, 2007; Nagaraja and Jayanna, 2013). Now these 1967 fused features increase the recognition percentage of MNSRS from 61.6-65.5% with 20 hidden neurons in MLFFN. And in the 3rd stage performance of MNSRS degrade from 65.5-61% by retraining with increase number of hidden neurons. For each 20, 40 and 60 hidden neurons we retrain 10 times each MLFFNN for selecting the best weight and bias values. Also we notice that Mean Square Error (MSE) and the Error percentage (E%) for training, validation and testing get increase in each stage. This occurs due over fitting MLFFNN.

MATERIALS AND METHODS

Ultilingual corpora collection: In this study, five native speakers Tamil, Hindi, Kannada, Malayalam and Telugu are considered for recording. These natives can speak six languages English, Tamil, Hindi, Kannada, Malayalam and Telugu. Multilingual text independent numbers (0-9) has taken for training and (10-19) for testing. These numbers are spoken by nine speakers in all six languages. Hence, the total native speakers in all five languages are 9×5 = 45 speakers. Each speaker speaks 10 numbers (0-9) in 6 languages. So the total number of samples per native speaker is 6×10 = 60 samples and total number of samples per native language is 9×60 = 540 samples. Finally, the total number of samples recorded for all natives are 540×5 = 2700 samples. Recording has done in canteen, hostel and class room and road in Hindustan University Padur Chennai using auto call recorder in spice mobile. During the recording voice signal is inferred by other signals.

Noise cancelation by ANFIS noise cancellation: In order to enhance multilingual native speaker on mobile phone we need cleaned voice signal in such a manner equal range of frequencies should be filter. To enhanced

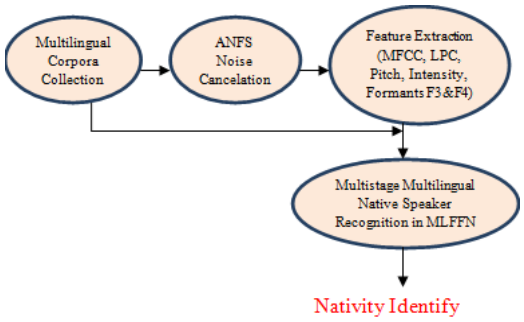

Fig. 1: Multistage MNSRS

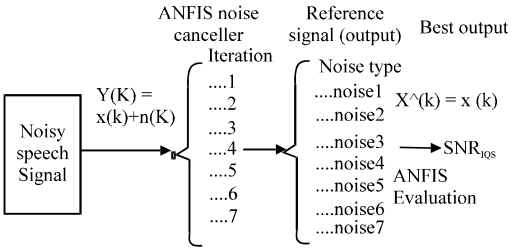

Fig. 2: ANFIS system for noise cancelation

MNSRS ANFIS system has used for active noise cancellation. In Fig. 2 complete ANFIS system we have presented. As we don't have any clean signal for reference signal only corrupted voice signal y(k) = x(k)+n(k) has taken for input. And different noise has taken as reference signal. For one corrupted voice signal we have taken 7 different type of noise (car, crowd, bike background, TV, fan, etc.). The 7 times (noise1 for iteration1 in ANFIS noise cancelation) as reference signal. We have calculated SNR and evaluated each ANFIS output and selected the best one $\hat{x}(k) = x(k)$ with highest SNR.

Feature extraction: Feature extraction is the key step to recognize MNSRS on mobile. In this study, three different types of features are extracted from noisy and clean speech. First auditory non linear 1950 MFCC features are extracted with Noise data and ANFIS cleaned data. For extracting 1950 MFCC 13 mel features has taken in each frame (total frames = 150). Other 13 vocal features Linear Predictive Coefficient (LPC) and 4 speech related prosody features (pitch, intensity, formants f3 and f4) are also extracted with ANFIS clean voice. LPC gives the estimation value of the current sample of a discrete signal as a linear combination of several previous samples. Prosody features give the detail about speech related properties.

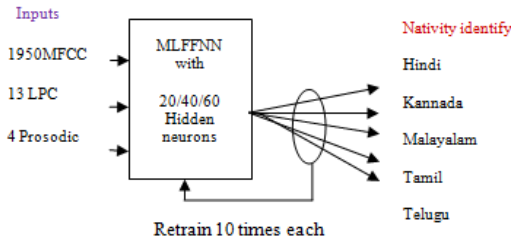

Fig. 3: MLFFNN structure with fused features

Finally, we fused 1950 MFCC, 13 LPC and 4 prosody features together (1967 fused features) to enhance the MNSRS.

Multilingual native speaker recognition system (MNSRS): MNSRS has done by MLFFNN in neural network. MLFFNN network has feed-forward multiple layers of computational units. Each neuron in one layer has directed connections to the neurons of the next layer. In this paper sigmoid function has used as an activation function. Error back-propagation used for calculating error in each no of iteration. Here, the output values are compared with target by calculating some predefined error-function using gradient descent method. Using this feedback, the algorithm adjusts the weights of each connection in order to reduce the value of the error function by some small amount. After some iteration, the network usually converge to some state where the error. MLFFNN applied in three stages

In the first stage, MLFFNN applied on noisy and ANFIS cleaned data with 1950 MFCC inputs value. In the second stage MLFFNN applied with 1967 fused features. In the third stage, Fig. 3 shows retrain the system 10 times for each 20, 40 and 60 hidden neurons in MLFFNN with 1967 fused features. In all three stages 75% data is taken for training, 15% for validation and 15% for testing the data. We have also calculated Mean Squared Error (MSE) and error percentage (E%).

MSE: Mean square error is an average squared difference between outputs and targets. MSE with low values are better, if it is zero means no error.

E percentage: Percent error indicates the fraction of samples which are misclassified. A value of E percentage is 0 means no misclassifications and 100 indicates maximum misclassifications.

RESULTS AND DISCUSSION

First stage enhancement with ANFIS cleaned data: After ANFIS noise cancelation we recognize nativity of speakers by using MLFFNN. We observe that with noisy data overall recognition is 44.3% and with ANFIS cleaned voice signal overall recognition increased by 53.7%. Complete detail is shown in Table 1. We notice that Hindi native recognized with higher accuracy 65.2% than other natives. The best performance before ANFIS is 0.12551 at 89 epochs and after ANFIS best performance increased by 0.1396 at 87 epochs. And also MSE%, E% get reduce after ANFIS.

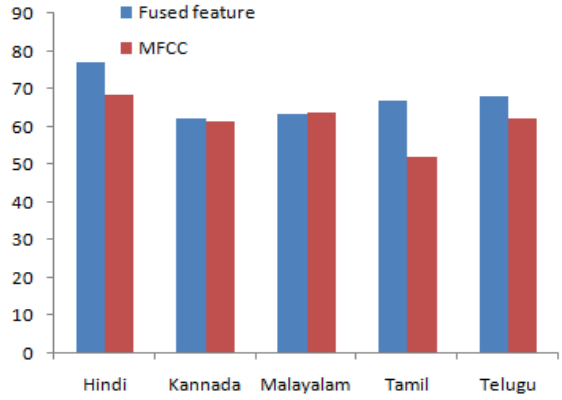

Fig. 4: Native recognition percent between fused features and MFCC alone

Table 1: First stage enhancement with ANFIS cleaned data

Variables	Before ANFSI (Noisy signal)	After ANFSI (cleaned signal)
MSE		
Training	0.1383	0.105
Validation	0.146	0.1293
Testing	0.149	0.1284
Error percentage		
Training	53.73	35.19
Validation	59.02	47.16
Testing	64.45308 (57)	50.37
Nativity recognatio (%)		
Hindi	308 (57)	352 (65.2)
Kannada	236 (43.7)	255 (47.2)
Malayalam	259 (48)	322 (59.6)
Tamil	251 (40.5)	242 (44.8)
Telugu	143 (26.5)	280 (51.9)

Perforance; 0.12551 at 89 epoch; 0.1396 at 87 epoch

Second stage enhancement by fusing the features: In the second stage, total 1967 features has taken as input by fusing of 1950 MFCC features with 13 LPC and 4 prosodic (pitch, intensity formant 3 and 4) in MLFFNN. We observed that overall recognition percentage increase from is 61.6-65.5% after fusing features. And also notice that Hindi speaker recognized mostly 76.9% with fused features. The best performance after fusing features is 0.12938 at 101 epochs. In Table 2, we have compare MSE% and E% with 1950 MFCC alone and with 1967 fused features. We notice that MSE and E percentage get reduce with reduce epochs for training, validation and testing data after fusing features.

In Fig. 4, we have shown the variation of recognition percentage for each native with fused features and MFCC alone.

Table 2: MSE and error percentage between fused features and MFCC alone

		MSE (%)			E (%)		
Features	Epochs	Tr	vld	Ts	Tr	vld	Ts
MFCC+LPC+Prosobic (fused 1967)	107	0.091	0.1201	0.1119	28.51	45.925	47.15
MFCC (1950)	178	0.105	0.1239	0.1264	35.18	47.160	50.37

Table 3: MSE and E percentage with 20, 40 and 60 hidden neurons

			MSE (%)			E (%)		
Hidden neurons	Trcyle	Epochs	Tr	vld	Ts	Tr	vld	Ts
20	4	176	0.0894	0.1221	0.1301	28.51	45.92	49.11
40	7	105	0.0998	0.1323	0.1314	32.6455	51.81	50.61
60	4	93	0.10332	0.1390	0.13966	35.87	53.35	51.85

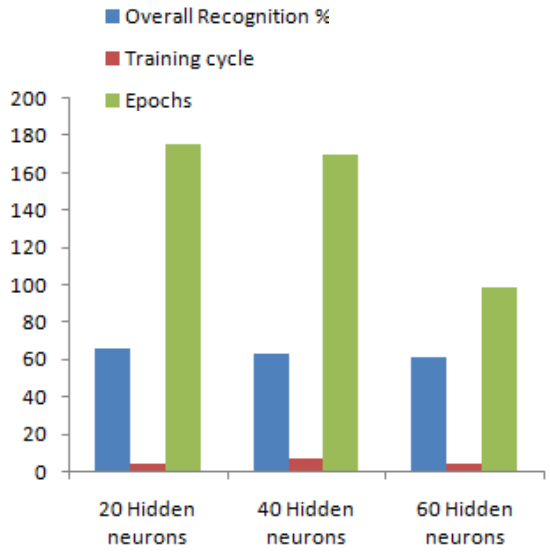

Fig. 5: Overall Recognition percentage, training cycle and Epochs in 20, 40 and 60 hidden neurons

Third stage MNSRS: In this stage, 1967 fused features have taken as input to MLFFNN. We retrain MLFNN for 10 times with 20, 40 and 60 hidden neurons. In Table 3, we have shown the MSE and error percentage with 20, 40 and 60 hidden neurons in various retraining cycle. We have shown overall recognition percentage, training cycle and epochs in Fig. 5. We notice that with 20 hidden neurons highest overall recognition percentage is 65.5 with 176 epochs in fourth training cycle and the best performance is .12215 at 170 epochs. For 40 hidden neurons highest overall recognition is 63.3% with 105 epochs in seventh training cycle and the best performance is 0.11231 at 99 epochs. For 60 hidden neurons highest overall recognition is 61 percentage with 105 epochs in fourth training cycle and the best performance is 0.1029 at 87 epochs.

Here an interesting fact, we notice that by retraining and increasing the number of hidden neurons overall native recognition fall down (65.5-61%). This is because of over fitting which degrade overall performance of MLFFNN.

CONCLUSION

In order to enhance MNSRS in first stage noise has removed from corrupted speech signal by ANFIS noise cancelation with increased SNR. In the second stage, we have fused auditory 1950 MFCC features with 13 LPC vocal features and 4 prosodic speech related features. That increase MNSRS from 61.6-65.5%. In the third stage, performance degrade due to over fitting by retraining MLFFNN with increase number of hidden neurons (20, 40 and 60). In future this research will extend to overcome over fitting and enhance MNSRS by using Convolutional Neural Network (CNN).

REFERENCES

Fredrickson, S.E. and L. Tarassenko, 1995. Text-independent speaker recognition using neural network techniques. Proceedings of the Fourth International Conference on Artificial Neural Networks, June 26-28, 1995, IEEE, UK, ISBN: 0-85296-641-5, pp: 13-18.

Hossain, M., M. Rahman, U.K. Prodhan and M. Khan, 2013. Implementation of back-propagation neural network for isolated bangla speech recognition. Int. J. Inf. Sci. Tech., Vol. 3.

Hosseinzadeh, D. and S. Krishnan, 2007. Combining vocal source and MFCC features for enhanced speaker recognition performance using GMMs. Proceedings of the International Conference on Acoustics, Speech and Signal Processing, April 15-20, 2007, Honolulu, HI., pp: 365-368.

Kelebekler, E. and M. Inal, 2006. White and Color Noise Cancellation of Speech Signal by Adaptive Filtering and Soft Computing Algorithms. In: Advanced In Artificial Intelligence. Sattar, A. and B.H. Kang (Eds.). Springer Berlin Heidelberg, Berlin, Germany, ISBN: 978-3-540-49787-5, pp: 970-975.

Martinez, J., H. Perez, E. Escamilla and M.M. Suzuki, 2012. Speaker recognition using Mel Frequency Cepstral Coefficients (MFCC) and Vector Quantization (VQ) techniques. Proceedings of the 22nd Annual International Conference on Electronics, Communications and Computers CONIELECOMP 2012, February 27-29, 2012, Yokohama National University, Cholula, Mexico, pp: 248-251.

Nagaraja, B.G. and H.S. Jayanna, 2013. Combination of features for multilingual speaker identification with the constraint of limited data. Int. J. Comput. Appl., Vol. 70.

Praveen, N. and T. Thomas, 2013. Text dependent speaker recognition using MFCC features and BPANN. Int. J. Comput. Appl., Vol. 74.

Suresh, L.R.D. and S. Sundaravadivelu, 2008. Real time adaptive nonlinear noise cancellation using fuzzy logic for optical wireless communication system with multiscttering channel. Eng. Lett., Vol. 13.