Abstract—This paper presents a distinctive phonetic features (DPFs) based phoneme recognition method by incorporating syllable language models (LMs). The method comprises three stages. The first stage extracts three DPF vectors of 15 dimensions each from local features (LFs) of an input speech signal using three multilayer neural networks (MLNs). The second stage incorporates an Inhibition/Enhancement (In/En) network to obtain more categorical DPF movement and decorrelates the DPF vectors using the Gram-Schmidt orthogonalization procedure. Then, the third stage embeds acoustic models (AMs) and LMs of syllable-based subwords to output more precise phoneme strings. From the experiments, it is observed that the proposed method provides a higher phoneme correct rate as well as a tremendous improvement of phoneme accuracy. Moreover, it shows higher phoneme recognition performance at fewer mixture components in hidden Markov models (HMMs).

Index Terms—distinctive phonetic features, language models, syllable based subword, inhibition/enhancement network, hidden Markov models

I. INTRODUCTION

Conventional automatic speech recognition (ASR) systems use stochastic pattern matching techniques in which word candidates are matched against word templates represented by hidden Markov models (HMMs). Although these techniques perform adequately in certain limited applications, they always reject a new vocabulary or a so-called out-of-vocabulary (OOV) word. Therefore, an accurate phonetic typewriter or a phoneme recognizer is expected to assist next-generation ASR systems for resolving this OOV-word problem [1, 2] via a short interaction (talk-back) by automatically adding the word into a word lexicon from the phoneme strings of an input utterance (see Fig. 1).

Though various distinctive phonetic features (DPFs) based methods, which can solve coarticulatory phenomena [3, 4], were proposed to obtain a more accurate phoneme recognizer [5, 6, 7, 8], they provide a higher phoneme correct rate (PCR) with poor phoneme accuracy rate (PAR). The reason for providing lower phoneme recognition accuracy is the violation of some phonotactic constraints at different places of the output phoneme string of an input utterance. Therefore, it is needed to incorporate syllable-based subword items,
TABLE I.

JAPANESE BALANCED DPF-SET FOR CLASSIFYING ATR PHONEMES

| DPFs | a | i | u | e | o | N | w | y | j | m | yk | yd | yb | yg | yn | yh | yr | yp | y | p | t | k | t | s | c | h | b | d | g | z | m | n | s | sh | h | f | r |
| vocalic | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| high | - | - | - | - | - | - | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| low | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| nil | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| anterior | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| back | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| nil | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| coronal | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| plasivce | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| affricative | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| continuant | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| voiced | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| unvoiced | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| nasal | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| semi-vowel | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |

which set some grammatical constraints at the recognition stage, for obtaining a phoneme recognizer with higher accuracy.

In this paper, we propose a more accurate phoneme recognition method by incorporating syllable language models (LMs), where the method comprises three stages. The first stage extracts three DPF vectors of 15 dimensions each from local features (LFs) of an input speech signal using three multilayer neural networks (MLNs). The second stage incorporates an Inhibition/Enhancement (In/En) network to obtain more categorical DPF movement and decorrelates the DPF vectors using the Gram-Schmidt (GS) algorithm. Then, the third stage embeds acoustic models (AMs) and LMs of syllables at the recognition stage of an HMM-based classifier. The objective of this study is to incorporate rule based (syllable constraints) LMs for resisting those phoneme strings that violates syllable rules and consequently, for obtaining a higher PAR.

The paper is organized as follows: Section II discusses the DPFs and Section III describes a hybrid neural network-based phoneme recognition method. Section IV explains the system configuration of the proposed phoneme recognition method. Experimental database and setup are provided in Section V, while experimental results are analyzed in Section VI. Finally, in Section VII, some conclusions are drawn.

II. DISTINCTIVE PHONETIC FEATURES

A phoneme can easily be identified by using its unique DPF set [9, 10]. The Japanese balanced DPF set [11, 12] for classifying phonemes have 15 elements which are vocalic, high, low, intermediate expression between high and low <nil>, anterior, back, intermediate expression between anterior and back <nil>, coronal, plosive, affricate, continuant, voiced, unvoiced, nasal and semivowel. Table I shows this balanced DPF set. Here, present and absent elements of the DPFs are indicated by “+” and “-” signs, respectively.

III. HYBRID NEURAL NETWORK-BASED PHONEME RECOGNITION METHOD

At first stage, this system [13] extracts a 45 dimensional DPF feature vector using a hybrid neural network (HNN) which consists of a RNN and an MLN, and then these features are decorrelated using the GS algorithm [11]. Here, the RNN is used for handling a longer context window [14]. Fig. 2 shows a block diagram of the system using HNN. Input acoustic vector is formed by taking preceding “t-3”-th and succeeding “t+3”-th frames together with the current t-th frame. Each input frame is formed by 25 LF values. The RNN generates 45 DPF values of which first 15 for the preceding, middle 15 for the current, and the rest for the succeeding frame. The MLN generates 45 DPF values and reduces misclassification of phoneme at phoneme boundaries by incorporating a seven frame context window (from t-3 to t+3) as input. At second stage, the system incorporates HMM-based AMs that are obtained using the orthogonalized feature vector of first stage, and LMs of syllable-based subword, which will be explained in Section IV.C, to output more precise phoneme strings.

IV. PROPOSED PHONEME RECOGNITION METHOD

Fig. 3 shows the proposed phoneme recognition method that comprises three stages. The first stage extracts a 45-dimensional DPF vector from the LFs of an input speech signal using three MLNs. The second stage incorporates In/En functionalities to obtain modified DPF patterns and decorrelates the DPF vector using the GS orthogonalization [11] before connecting it with an HMM-based classifier.

A. DPF Extractor

In this method, three MLNs instead of two MLNs [5] are used to construct the DPF extractor. The first MLN,
MLN_{LF,DPF}, outputs DPFs [11, 12] for the inputted acoustic features, LFs [15], while the second MLN, MLN_{cntxt}, reduces misclassification at phoneme boundaries by taking seven frame context (from t-3 to t+3) as input, and the third MLN, MLN_{Dyn}, restricts the DPF dynamics by incorporating dynamic parameters (ΔDPF and ΔΔDPF) into its input. Here, the MLN_{LF,DPF}, which is trained using the standard back-propagation learning algorithm, has two hidden layers of 256 and 96 units, respectively and takes three input vectors (t-3, t, t+3) of LFs of 25 dimensions each. The 45-dimensional context-dependent DPF vector provided by the MLN_{LF,DPF} at time t is appended into the MLN_{cntxt}, which consists of five layers including three hidden layers of 90, 180, and 90 units, respectively, and generates a 45-dimensional DPF vector with a small number of errors at phoneme boundaries by incorporating context window of particular size. This 45-dimensional DPF vector and its corresponding ΔDPF and ΔΔDPF vectors calculated by three-point linear regression (LR) are appended into the subsequent MLN_{Dyn}, which consists of four layers including two hidden layers of 300 and 100 units, respectively, and outputs a 45-dimensional DPF vector with reduced fluctuations and dynamics. Both the MLN_{cntxt} and MLN_{Dyn} are also trained using the standard back-propagation algorithm. Here, for the MLN_{LF,DPF} and MLN_{cntxt}, the context window size is selected <t-3, t, t+3> instead of <t-3, t-2, t-1, t, t+1, t+2, t+3> because approximately same performances are obtained by the both and reduced computational cost is required for the context window, <t-3, t, t+3>. On the other hand, the specified unit of each hidden layer for each MLN is selected by tuning process.

B. Inhibition/Enhancement Network

The DPF extractor provides 45 DPF patterns (15 for preceding context, 15 for current context, and 15 for following context) for each input vector along time axis. These patterns may not match with the pattern of input phoneme string for some phonemes and consequently, some phonemes are incorrectly recognized by the HMM classifier. This phoneme misclassification sometimes occurs when the values of DPF peaks and DPF dips are closer to each other. Therefore, a mechanism, which is called In/En network [5, 6, 7], is needed to obtain clearly

Figure 2. Hybrid Neural Network based Phoneme Recognition Method.

Figure 3. Proposed Phoneme Recognition Method.
separable DPF peaks and dips. An algorithm for this network is given below:

**Step1:** For each element of the DPF vectors, find the acceleration ($\Delta \Delta$) parameters by using three-point LR.

**Step2:** Check whether ($\Delta \Delta$) is positive (concave pattern) or negative (convex pattern) or zero (steady state).

**Step3:** Calculate $f(\Delta \Delta)$

\[ f(\Delta \Delta) = \frac{c_1}{1 + (c_1 - 1) e^{b \Delta \Delta}} \]  

if pattern is convex,

\[ f(\Delta \Delta) = \frac{2(1 - c_2)}{1 + e^{b \Delta \Delta}} \]  

if pattern is concave,

\[ f(\Delta \Delta) = 1.0 \]  

if steady state,

**Step4:** Find modified DPF patterns by multiplying the DPF patterns with $f(\Delta \Delta)$.

**C. Syllable-Based Language Model**

The phoneme classification inaccuracy at the acoustic phonetic level is a major weakness in most speech recognition systems. However, the inaccuracy often violates phonotactic constraints at the acoustic phonetic level. Usually, three types of phoneme classification errors occur: (i) phoneme insertions, (ii) phoneme deletions, and (iii) phoneme substitutions. A better performance is expected if a language model is adopted in a recognition system for post-processing phoneme estimates and for making corrections with regard to the grammatical constraints. Words, syllables, and phonological rules, etc. of a language provide some constraints on a phoneme sequence generation, but a language has a large number of words and knowledge-based rules. Besides, since the number of syllables of Japanese language is limited, they can be used to enhance the phoneme recognition performance. For an example, a phoneme recognizer generates a phoneme strings, /k/ /a/ /m/ /a/ /s/ /u/ for an input speech /kakemasu/, which means “write” in English language. The actual phoneme strings for the utterance is /k/ /a/ /m/ /a/ /s/ /u/. Therefore, there occur two insertions by the phoneme recognizer. After adopting LMs that support only CV, CVV, CV1V2, CVN, and CVQ syllable structures, we can correct the phoneme strings, /k/ /a/ /m/ /a/ /s/ /u/ by resisting the generation of strings, /k/ /a/ /m/ and /k/ /a/ /s/ /u/ incorrectly.

**V. EXPERIMENTS**

**A. Speech Databases**

If we use same data set for all classifiers used in the experiments, there are some possibilities of biasness for the test data to the training data at each stage. Because of this, open test for each stage classifier (neural network-based and HMM-based) is done by using different training and test data sets. The following five clean data sets are used in our experiments.

**D1. Training data set for MLN**

A subset of the Acoustic Society of Japan (ASJ) Continuous Speech Database [16] comprising 4503 sentences uttered by 30 different male speakers (16 kHz, 16 bit) is used.

**D2. Training data set for MLNcntxt**

This data set contains 5000 sentences that are taken from Japanese Newspaper Article Sentences (JNAS) [17] Continuous Speech Database; the sentences have been uttered by 33 different male speakers (16 kHz, 16 bit).

**D3. Training data set for MLNDyn**

This data set contains 5000 JNAS [17] sentences uttered by 33 different male speakers (16 kHz, 16 bit). Speakers of this data set are different from the D2 data set.

**D4. Training data set for HMM classifier**

This data set takes 5000 JNAS [17] sentences uttered by 33 different male speakers (16 kHz, 16 bit). Speakers of this data set are different from the D2 and D3 data set.

**D5. Test data set**

This test data set comprises 2379 JNAS [17] sentences uttered by 16 different male speakers (16 kHz, 16 bit).

**B. Experimental Setup**

The frame length and frame rate (frame shift between two consecutive frames) are set to 25 ms and 10 ms, respectively, to obtain acoustic features from an input speech signal. LFs are a 25-dimensional vector consisting of 12 delta coefficients along time axis, 12 delta coefficients along frequency axis, and delta coefficient of log power of a raw speech signal [15].

Since our goal is to design a more accurate phoneme recognizer, PCR and PAR for D5 data set are evaluated using an HMM-based classifier. The D4 data set is used for the proposed method to design 38 Japanese monophone HMMs with five states, three loops, and left-to-right models. Input features for the classifier are DPFs. On the other hand, in the HNN-based phoneme recognition method, the D1 data set is used to design the HMM and orthogonalized DPFs are used as input features. In the HMMs, the output probabilities are represented in the form of Gaussian mixtures, and diagonal matrices are used. The number of components per mixture is set to 1, 2, 4, 8, and 16, respectively.

In our experiments of the three-MLN and HNN, the non-linear function is a sigmoid from 0 to 1 ($1/(1+\exp(-x))$) for the hidden and output layers.

For the In/En network, the value of the enhancement coefficient, C1, is set to 4.0 after evaluating the proposed method, DPF (3-MLN+In/En+GS, dim:45), for different values of C1, such as 2, 4, and 6, and the value of the steepness coefficient, $\beta$, is set to 80. The value of inhibitory coefficient, C2, is fixed to 0.25 after observing the DPF data patterns to keep the values of $f(\Delta \Delta)$ between 0.25 and 1.0.

For the experiments incorporating LMs, the Julius 3.5.3 version [18] is used to embed acoustic models and trigram subword models (short and long syllables) as
LMs. It is a two-pass system, where the first pass is a bigram and the second one is a trigram to set more constraints on the output phoneme sequence. Syllable structures are CV, CVV, CV1V2, CVN, and CVQ. Weight/Insertion penalties are 5.0/-1.0 and 6.0/00 for the first and second passes, respectively.

To investigate the effect of syllable-based subword LM, we have designed the following phoneme recognition tests, where the DPF extractor inputs a 45-dimensional orthogonalized DPF vector to the HMM.

(a) 3-MLN+In/En+GS [7]
(b) 3-MLN+In/En+GS+LM [Proposed]

For further investigation, we have designed the following experiments for testing the phoneme recognition performance with and without incorporating LM.

(c) Hybrid+GS
(d) Hybrid+GS+LM [13]

To compare the methods, Huda [5] and the proposed one, we have designed the following two experiments for evaluating PCR and PA.

(e) 2-MLN+In/En+GS [5]
(f) 3-MLN+In/En+GS+LM [Proposed]

For comparing context window size, <t-3, t, t+3> for method (e) with <t-3, t-2, t-1, t, t+1, t+2, t+3> for method (f), we have evaluated PCR and PA by designing the following two experiments.

(f) Cont.2-MLN+In/En+GS
(e) 2-MLN+In/En+GS [5].

VI. EXPERIMENTAL RESULTS AND DISCUSSION

Fig. 4 shows the PCR of the proposed method before and after adding syllable-based subword LM. For all the mixture components investigated, the method with LM provides higher PCR. For example, at mixture component 16, the method without LM shows 84.50% PCRs, while the corresponding value for the method with LM is 87.40%.

On the other hand, the method with LM outperformed the method without LM for phoneme accuracy, which is shown in Fig. 5. By using the mixture components 1 and 2, the syllable knowledge leads to correction of the highest number of phonemes and shows the highest level improvement. Again, the addition of syllable knowledge shows the highest accuracy (80.54%) at mixture component 16. Therefore, language constraints resist misclassification of phonemes and consequently, provide higher phoneme recognition performance in a clean acoustic environment.

Again, Figs. 6 and 7 depicted the PCR and PAR, respectively, for the HNN-based method with and without incorporating syllable-based LM. Incorporation of LM improves the PCR and PAR for all the mixture components investigated. It is observed from the Fig. 6 that the language model improves the PCR most (6.92%) at mixture component eight, while the PAR is improved by 6.98%, which is shown in Fig. 7.

It is claimed that syllable-based subword increases PCR and PAR, where improvement of PAR is tremendous. The reason for improving PAR is that syllables setup some rules on phoneme generation. These rules resist some phonemes which violates the constraints set by syllables.
Segmentation for a clean /jiNkoese/ utterance is shown in Figs. 8 and 9 for a balanced-DPF set [11] using an HNN and 3-MLN, respectively. In both the figures, "nasal", "nil(high/low)", and "high" of phoneme /N/, and "unvoiced", "coronal", and "anterior" of phoneme /s/ are denoted by (1), (2), (3), (4), (5), and (6), respectively. After observing these marked places, we can say that the 3-MLN exhibits more precise segmentation (less deviation from ideal boundary) than the HNN, reduces some fluctuations and provides more smoothed DPF curves, and hence, it misclassifies fewer phonemes. Consequently, the 3-MLN based proposed method provides higher phoneme recognition performance than the HNN based method.

Tables II and III show PCR and PAR, respectively for the methods, [5] and proposed one in this study. From both the tables, it is observed that the incorporation of one additional MLN and LMs increases PCR and PAR for all mixture components investigated, where PAR improvement is tremendous than PCR. For an example, at mixture component 16, approximately 4.00% improvement of PCR is observed (see Table II), while the corresponding value of PAR is approximately 14.00% (see Table III).

Again, Tables IV and V show PCR and PAR, respectively for the methods, Cont.2-MLN+In/En+GS with continuous <t-3, t-2, t-1, t, t+1, t+2, t+3> and [5] with discrete <t-3, t, t+3>. It is shown from both the tables that the methods, Cont.2-MLN+In/En+GS with continuous seven frames and [5] with discrete seven frames provide almost same performances for all higher mixture components investigated. Moreover, higher dimensionality in the method with continuous seven frames requires more computation in the training process of neural networks and consequently, slows down both the training and test phase. Therefore, discrete <t-3, t, t+3> is selected for experiments.

Table VI shows the effect of different training datasets, for the same test data (D5), on phoneme recognition performance. From the table, it is observed that the methods, with same training data set for all classifiers (first data row of the table) and with different training data set for each classifier (second data row of the table), show almost same performances for all the mixture components investigated.

**TABLE II.**

<table>
<thead>
<tr>
<th>Methods</th>
<th>Phoneme Correct Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 mix</td>
</tr>
<tr>
<td>2-MLN+In/En+GS [5]</td>
<td>82.18</td>
</tr>
<tr>
<td>Proposed</td>
<td>84.10</td>
</tr>
</tbody>
</table>

**TABLE III.**

<table>
<thead>
<tr>
<th>Methods</th>
<th>Phoneme Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 mix</td>
</tr>
<tr>
<td>2-MLN+In/En+GS [5]</td>
<td>56.53</td>
</tr>
<tr>
<td>Proposed</td>
<td>76.55</td>
</tr>
</tbody>
</table>
TABLE IV.
PHONEME CORRECT RATE FOR THE METHODS, CONT.2-
MLN+In/En+GS with continuous <T-3, T-2, T-1, T, T+1, T+2, T+3>
and [5] with discrete <T-3, T, T+3>

<table>
<thead>
<tr>
<th>Methods</th>
<th>Phoneme Correct Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 mix</td>
</tr>
<tr>
<td>Cont.2-MLN+In/En+GS</td>
<td>82.84</td>
</tr>
<tr>
<td>2-MLN+In/En+GS[5]</td>
<td>82.18</td>
</tr>
</tbody>
</table>

TABLE V.
PHONEME ACCURACY RATE FOR THE METHODS, CONT.2-
MLN+In/En+GS with continuous <T-3, T-2, T-1, T, T+1, T+2, T+3>
and [5] with discrete <T-3, T, T+3>

<table>
<thead>
<tr>
<th>Methods</th>
<th>Phoneme Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 mix</td>
</tr>
<tr>
<td>Cont.2-MLN+In/En+GS</td>
<td>60.47</td>
</tr>
<tr>
<td>2-MLN+In/En+GS[5]</td>
<td>56.53</td>
</tr>
</tbody>
</table>

TABLE VI.
EFFECTS OF TRAINING DATASETS ON PHONEME RECOGNITION
PERFORMANCES

<table>
<thead>
<tr>
<th>Training and Test Data Set for each Classifier</th>
<th>Phoneme Correct Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train: MLN+9D1, MLN+9D1, HMM+9D1 Test: D5</td>
<td>81.76</td>
</tr>
<tr>
<td>Train: MLN+9D1, MLN+9D1, HMM+9D4 Test: D5</td>
<td>81.59</td>
</tr>
</tbody>
</table>

VI. CONCLUSION

In this paper, a DPF-based phoneme recognition method is presented by incorporating Japanese syllable-based LMs. From the experiments using JNAS database, the following conclusions are drawn.

i) The proposed method with LMs increases phoneme correct rate by approximately 2% to 3% in comparison with the method [7] for all the mixture components investigated (see Fig. 4), while this improvement value is 2% to 4% compared to the method [5] (see Table II).

ii) It improves phoneme recognition accuracy by 9% to 18% in comparison with the method [7] (see Fig. 5). On the other hand, 14% to 20% improvement is observed by the proposed method compared to the method [5] (see Table III).

iii) It requires fewer mixture components to obtain a higher recognition performance.

In future research, an evaluation of Bengali phoneme recognition accuracy will be tried using the method proposed in this paper.

REFERENCES


Manoj Banik was born in Habiganj, Bangladesh in 1972. He completed his B.Sc. in Computer Science and Engineering Degree from BUET, Dhaka, Bangladesh. Now, he is working as an Assistant Professor in the Department of Computer Science and Engineering of Ahsanullah University of Science and Technology, Dhaka, Bangladesh. His research interests include Speech Recognition, Phonetics and Algorithms. He is a member of Bangladesh Engineers Institute (IEB).

Ghulam Muhammad was born in Rajshahi, Bangladesh in 1973. He received his B. Sc. in Computer Science and Engineering degree from Bangladesh University of Engineering & Technology (BUET), Dhaka in 1997. He also completed his M.E and Ph. D from the Department of Electronics and Information Engineering, Toyohashi University of Technology, Aichi, Japan in 2003 and 2006, respectively. Now, he is working as an Assistant Professor in King Saud University, Riyadh, Saudi Arabia. His research interest includes Automatic Speech Recognition and human-computer interface. He is a member of IEEE.

Mashud Kabir was born in Narayanganj, Bangladesh in 1976. He has completed his Bachelor of Science (BSc) in Electrical & Electronic Engineering from Bangladesh University of Engineering & Technology (BUET) in 2000. He earned his Master of Science (MSc) in Communication Engineering from University of Stuttgart, Germany in 2003. He worked at Mercedes-Benz Technology Center, Germany from 2003 to 2005 as a PhD student. He has work-experience in research & development projects of BMW, Land-Rover and Audi in automotive infotainment network and electronic area for more than five years. The author achieved his PhD degree from the University of Tuebingen, Germany in 2008. He is working as an Assistant Professor in United International University, Dhaka, Bangladesh.

Bernd J. Kröger received his Master's degree in Physics, University of Münster in 1985. He also received his PhD degree in Phonetics, University of Cologne, Germany in 1989. Now, he is working as a senior scientific staff member at Department of Phoniatrics, Pedaudiology, and Communication Disorders, UKA and RWTH Aachen University. His research fields are Articulatory feature–based Automatic Speech Recognition and Speech Synthesis.