Voice Morphing Based on Spectral Features and Prosodic Modification

A. Qavi and S. A. Khan
Department of Electrical and Computer Engineering, Center for Advanced Studies in Engineering (CASE), Islamabad, Pakistan
Email: aqavi_paracha@yahoo.com, shoab@case.edu.pk

K. Basir
Department of Computer Engineering, College of Electrical and Mechanical Engineering (CEME), National University of Science and Technology (NUST), Islamabad, Pakistan
Email: kashifbasir@yahoo.com

***Abstract*—This paper is aimed at morphing the speech uttered by a source speaker in a manner that it seems to be spoken by another target speaker – a new identity is given while preserving the original content. The proposed method transforms the vocal tract parameters and glottal excitation of the source speaker into target speaker's acoustic characteristics. It relates to the development of appropriate vocal tract models that can capture information specific to the speaker and estimate the model parameters that closely relate to the model of the target speaker. It detects the pitch, separates the glottal excitation and vocal tract spectral features. The glottal excitation of the source is taken, voice/un-voice decision is made, the prosody information is found, PSOLA is used to modify the pitch, the spectral features are found, and finally speech is modified using target spectral features and prosody. The subjective experiment shows that the proposed method improves the quality of conversion and contains the original vocal and glottal characteristics of the target speaker.**

***Index Terms*—voice morphing, spectral features, resampling, voice/un-voice activity detection, windowing, prosody, PSOLA**

I. Introduction

In this paper, we morph the voice of the source speaker to that of target based on spectral features and prosodic information in the speech. Here, voice morphing targets the non-linguistic features of speech signals, such as voice identity, voice quality and pitch. It is yet a growing field with quite attractive applications, giving enough space for new research and development. This speech handling gets even more complex when one has to handle various accents and abstract rules of communication – the language.

One has to understand the voice production and perception mechanism first to efficiently implement the voice morphing system. In the start, one of the first methods for voice morphing was proposed by R. J. McAulay and Quatiery based on Sinusoidal modeling [1]. It provided very basic transformation of the voice, but won't scale for large speech signal. K. Tanaka and M. Abe improved upon Sinusoidal to introduce a new pitch modification algorithm along with conversion of spectrum [2]. Next, Levent M. Arslan and David Talkin introduced a new Spectral Transformation Algorithm Using Segmental Codebooks (STASC) where they presented two new methods to perform conversion of vocal tract and glottal excitation characteristics [3]. In addition, they also proposed time-scale and pitch-scale modification algorithm. Then, Allam Mousa proposed a new conversion system using pitch-shifting by time-stretching with PSOLA for Arabic speech [4]. Many authors proposed different methods like Linear Predictive Coefficients (LPC), Formant Frequencies (FF), residual-excited LPC (RELP), Line Spectral Frequencies (LSF), Mel Frequency Cepstrum Coefficient (MFCC), harmonic pulse-noice model parameters or mix of time and frequency domain methods to change the spectral features, pitch and duration [5]. Although most of the above methods performed transformation, which was better in one way or the other, but they did not consider other important characteristics of human speech, like nasal cavity, pitch and other non-linguistic features [6]. Most of these methods neglected fine details about excitation component and formant features, which produced muffled effect in the morphed speech. So for the enhancement of the speech, another approach called STRAIGHT was proposed, but this method requires enormous computation so is not suitable for real-time applications [7], [8]. Nevertheless, there is no one unique best method for features extraction. Each method has its own pros and cons. D. Erroet all used MFCC and F0 to extract features and used HNM for voice conversion [9]. Both Peramanallu and Rajvi shah have separately used LSF and LPC for voice conversion respectively [10]. Yet our evaluation section shows that our results outperform the results attained either by using MFCC, Complex cepstrum or other LPC approaches. Keeping in view the trade-off of computation power required and best vocal tract feature extraction methods, we proposed a method

Manuscript received July 28, 2014; revised October 23, 2014.

doi: 10.12720/ijsps.3.2.123-127

where the spectral features of both the source and target speakers are taken using Linear Prediction, along with prosodic information, considering pitch, nasal cavity, lips radiation and other fine details in excitation. The previously introduced methods lacked one feature or the other. The voice is finally transformed based on all the features of target speaker. Although this method gives quite good results, it's yet dependent on text, i.e., same text has to be spoken by the source and target speakers in order for it to transform the speech.

II. Speech Production Mechanism and the Model

Human speech production mechanism is non-linear phenomena and requires deep study to understand it well. The air is pumped from the lungs, passes through the vocal tract and lips to produce the final speech. Fig. 1 models the human speech production mechanism.

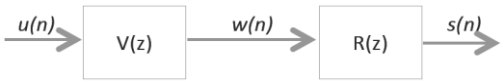

Figure 1. Human speech production model

In Fig. 1, the source signal is represented by u(n), and is modeled by the transfer function of human vocal tract V(z), which produces w(n) – the lips velocity. Then R(z), the lips radiation filter further changes the signal spectrum. The term $1 - z$ can be used to approximate the delay term.

$$V(z) = \frac{Gz^{-\frac{p}{2}}}{1-\sum_{j=1}^{P} a_j z^{-j}} = \frac{Gz^{-\frac{p}{2}}}{A(z)} \triangleq G/A(z) \qquad (1)$$

An all-pole filter can be used to approximate the transfer function of human vocal tract. Refer to the (1) where $\frac{1}{A(z)}$ represents vocal tract transfer function

III. The Voice Morphing Model

In this section, we describe the voice morphing model we used. Fig. 2 shows the various stages of the source and target speech decomposition and then synthesizing to produce the morphed speech signal.

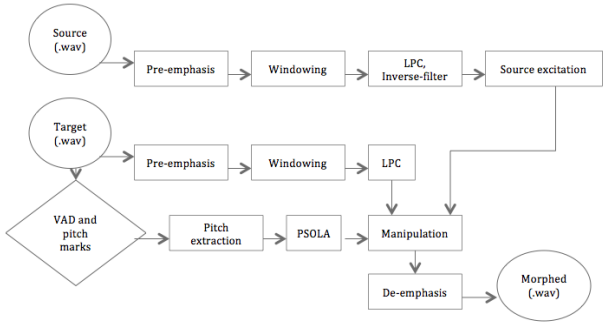

Figure 2. Block diagram of the speech morphing system

A. Preemphasis/Deemphasis

Prior to the analysis, the speech signal is passed through a pre-emphasis filter in order to reduce the dynamic range of speech spectrum. The pre-emphasis is applied to the input signal before the LPC analysis, while the de-emphasis is applied during reconstruction following the LPC analysis in order to revert the effect. Fig. 3 shows the pre-emphasized signal. Pre-Emphasis and de-emphasis are needed because in human speech spectrum, the energy lowers as the frequency gets high.

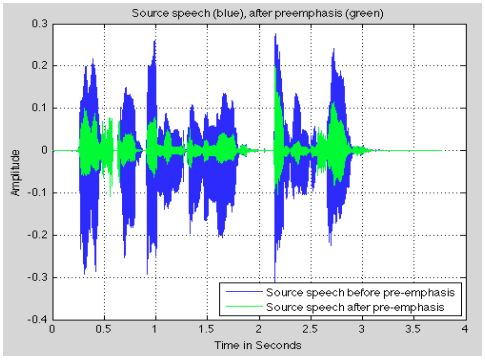

Figure 3. The actual speech (blue), the speech after pre-emphasis (green)

B. Windowing

The pre-emphasized speech is then segmented into short-term frames for analysis using a Hamming window. The experiments show that the human pitch does not go below 50Hz, which approximates to 20ms duration. So we take the frame size of 30ms to cover at least 2 pitch periods. We used Hamming window because of its tapered frequency response. It lowers the effect of discontinuities at the boundaries of each analysis frame. Fig. 4 shows the effect of hamming window. Hamming window is described by the following equation.

$$w(n) = 0.54 - 0.46\cos\left(2\pi\frac{n}{N-1}\right) \qquad (2)$$

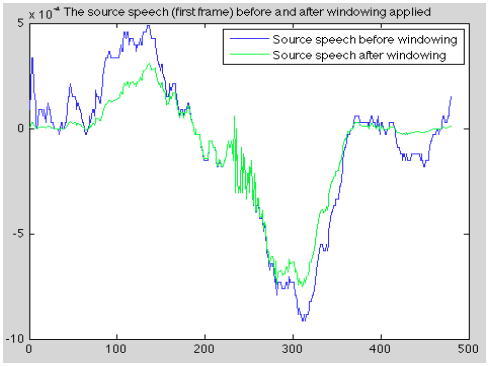

Figure 4. The pre-emphasized speech (blue), the speech after windowing applied (green).

C. Voice/Un-Voice Decision

In this step, the speech frames are first isolated as voiced/un-voiced so that pitch extraction (explained in next section) is applied to the voice frames only. Fig. 5 describes the steps involved in making V/U decision of frames.

Figure 5. Block diagram of voice/un-voice decision

D. Spectral Features Modification

After the analysis phase, the Linear Predictive Coefficients are determined for both the source and target speakers. These features are used to model the vocal tract filter of the target speaker, and an inverse filter is used to extract the excitation component of the source speaker. Now this excitation component of source is applied to the all-pole vocal tract filter of target, which shapes the spectrum of source speaker's excitation. Thus by this filtering operation, spectral shaping of source signal is achieved, to which the converted pitch is applied to finally achieve the morphed voice.

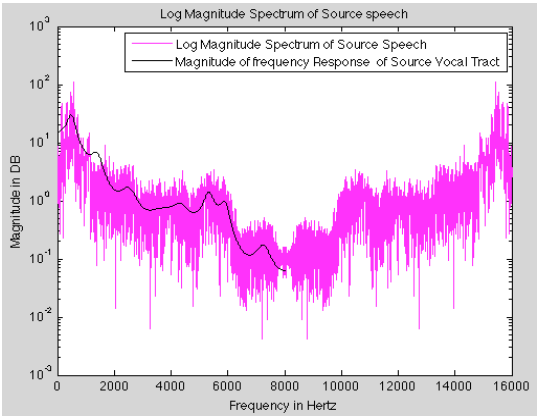

Figure 6. Magnitude of frequency response of the source vocal tract

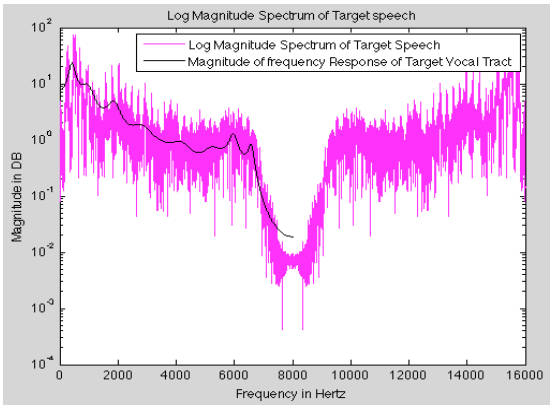

Figure 7. Magnitude of frequency response of the target vocal tract

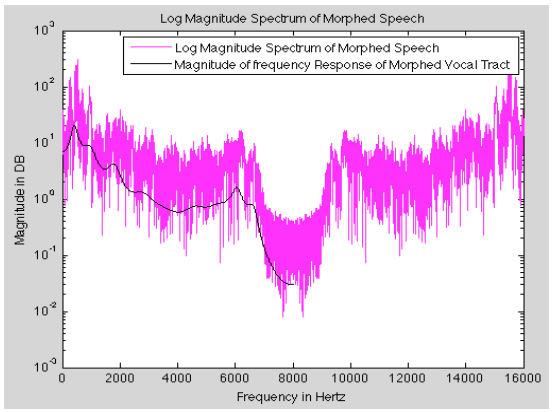

Figure 8. Magnitude of frequency response of the morphed speech

Fig. 6 and Fig. 7 show the magnitude of frequency response of the source and target speakers' vocal tract. Fig. 8 shows the magnitude of frequency response of the morphed speech, which closely resembles the target speaker's vocal tract frequency response

E. The Excitation Part

Even though the human vocal tract contains most part of the speech, the rest of the information is yet contained in residual signal, which is quite tricky and challenging to model. With the inverse filtering, we obtain an estimate of the glottal filter. In the modification process, we take the excitation of the source speech, and apply the target speaker's vocal tract features. Fig. 9 shows the extraction of excitation from the source speech frame.

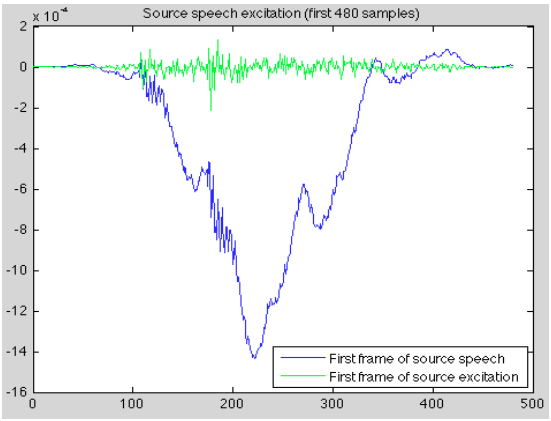

Figure 9. The excitation extracted from the source speech frame

F. The Pitch Extraction and PSOLA

The excitation component is first modeled based on detection of voiced/unvoiced signal, after which an auto-correlation method is used to compute pitch. For each frame, voice detection flag is set. The pitch value is determined only for voiced component and the pulse train of determined pitch period is generated, whereas for the unvoiced components white Gaussian noise is generated. The unvoiced components are much noise-like and have very less energy as compared to the voiced components (which have high energy).

PSOLA (Pitch-Synchronous OverLap Add) is then used to convert the pitch from source to target speech. Fig. 10 and Fig. 11 are produced using Wavesurfer, of which Fig. 10 shows the pitch of both the source and target speakers, and Fig. 11 shows the pitch of morphed speech, which closely resembles the pitch of target speaker.

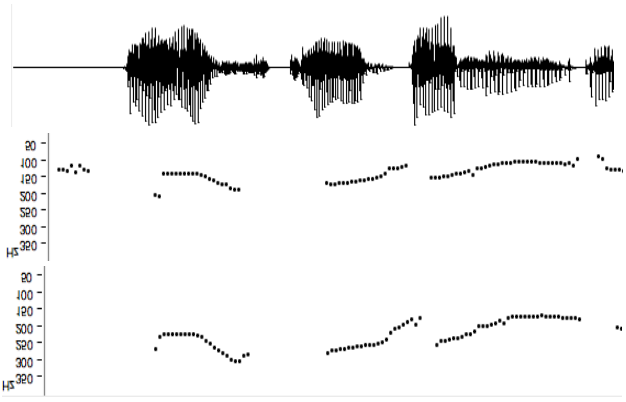

Figure 10. Pitch of the source (above) and target (below)

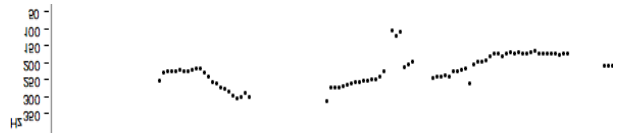

Figure 11. Pitch of the morphed speech

IV. Applications

Voice morphing has numerous applications in the industry, such as dubbing movies by preserving the original actors' voices, and security where you hide the actual identity of the speaker. In commercial fields, it can be applied in toys for entertainment. In Interactive Voice Response systems (IVR), one can listen to email messages on phone via mail readers. Text-independent voice morphing can be used to preserve the voices of great singers for many years, reproducing the appropriate voices in many applications without the original speaker being present such as in broadcasting. An ordinary person's speech can be converted into a famous singer's speech for entertainment. Virtual voice generation is another application in 3D movies. It is closely related to speech synthesis (converting text into spoken language) that has many applications esp. for deaf and blind.

Another application is for the people suffering from Laryngectomy [11] – the surgical removal of the larynx (particularly performed in case of laryngeal cancer). In this case, extracting the source's vocal coefficients and converting them into the target speaker's vocal characteristics through training achieves voice morphing. There are different approaches to training, such as Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Artificial Neural Networks (ANN), Radial Basis Functions (RBF), and others. Besides vocal characteristics of the source and target speakers, the pitch information is also transformed. With this people suffering from laryngeal cancer may be able to speak.

V. Performance Evaluation and Results

The morphing system is tested using speech data from voice databases – TIMIT database, RWTH Aachen, and Cambridge University. The corpus contains many different recorded speeches from single word to phrase and sentences. In our experiments, four different voice conversion tasks were investigated: male-to-male, male-to-female, female-to-male, and female-to-female conversion.

Fig. 12 shows the modeling of human vocal tract of source speaker (blue), target speaker (cyan) and the morphed speech (green). It shows how closely the target speaker's vocal tract filter resembles the morphed speech filter.

Fig. 13, Fig. 14 and Fig. 15 show the results of the source and target speeches along with the morphed speech. Two speakers speak approximately 3seconds sentence "Don't ask me to carry an oily rag like that", and then morphing is applied. Fig. 15 shows the morphed speech signal, which closely resembles the target speech signal.

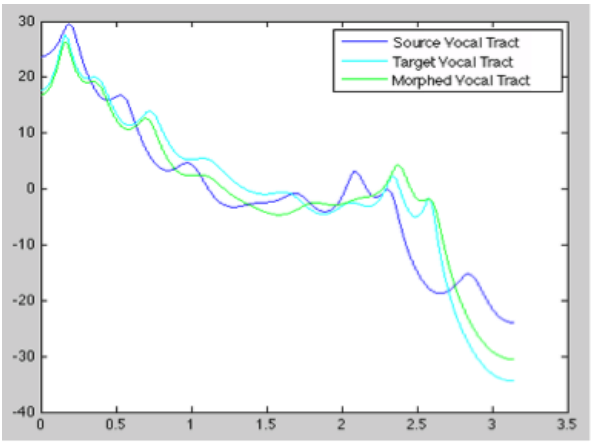

Figure 12. The human vocal tract model of source (blue), target (cyan), and morphed (green)

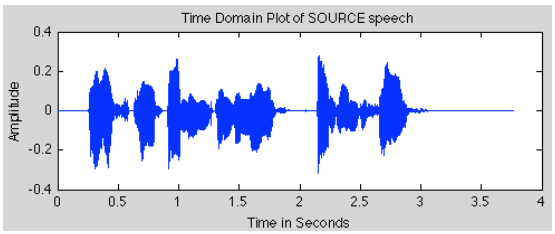

Figure 13. Source speech uttering the sentence "Don't ask me to carry an oil rag like that"

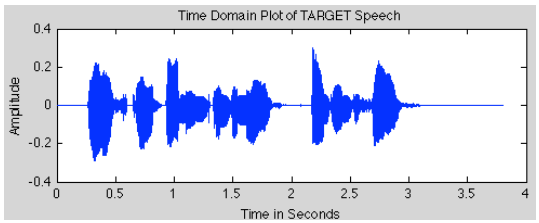

Figure 14. Target speech uttering the same sentence "Don't ask me to carry an oil rag like that"

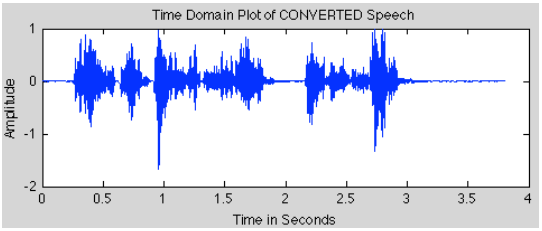

Figure 15. Morphed speech uttering the same sentence "Don't ask me to carry an oil rag like that"

In order to evaluate the performance of the system, ABX-style test was performed, which is common practice for voice morphing evaluation tests [12], [13]. A panel of 20 listeners was asked to evaluate the performance of the system. They were presented with all the three sounds A, B and X, and were asked to tell whether the utterance X sounds close to the A or B, where A and B were original source and target speeches and X the morphed speech. In total, 16 random utterances were tested, which consisted of 4 tests for each male-to-male, female-to-female, female-to-male and male-to-female. Table I depicts the results of the test in percentage.

TABLE I. THE ABX STYLE TEST RESULTS FOR THE EVALUATION OF THE SYSTEM

Experiments	Correct percentage (%)
Male-Female	90%
Female-Female	70.25%
Male-Male	82.5%
Female-Male	85%

VI. CONCLUSION AND FUTURE WORK

Our system can morph the two voices when both the speakers utter the same sentence, phrase or a word. This system is fairly good at impersonating a certain target speaker. From the results, it can be implied that the system is better at converting both male and female speakers' voices, though a small rigging effect is noticed.

To make the system more robust, efficient and text-independent, we extend the algorithm to introduce a learning method so that the morphing system can be trained with target speaker's vocal parameters, which can finally be used to transform any text spoken by the source speaker. Based on studies, we've found the best results can be obtained using Radial Basis Function Networks [8], [14], [15]. In addition, the system can be made even more efficient by introducing time alignment of the speeches before training phase.

REFERENCES

[1] R. J. McAulay and T. F. Quatiery, "Sinusoidal coding," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds., New York: Elsevier Science Inc., 1995.

[2] K. Tanaka and M. Abe, "A new fundamental frequency modification algorithm with transformation of spectrum envelope according to f0," in *Proc. ICASSP*, Munich, 1997, pp. 952-954.

[3] L. M. Arslan, "Speaker transformation algorithm using segmental codebooks (STASC)," *Journal Speech Communication*, vol. 28, no. 3, pp. 211-226, Jul. 1999.

[4] A. Mousa, "Voice conversion using pitch shifting algorithm by time stretching with PSOLA and re-sampling," *Journal of Electrical Engineering*, vol. 61, no. 1, pp. 57-61, 2011.

[5] N. Davel, "Feature extraction methods LPC, PLP and MFCC in speech recognition," *International Journal for Advance Research in Engineering and Technology*, vol. 1, no. VI, Jul. 2011.

[6] X. Lu and J. Dang, "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification," *Speech Communication*, vol. 50, no. 4, pp. 312-322, 2008.

[7] G. Xu, Q. Zou, D. Zhao, and D. Huang, "Straight model for voice conversion based on acoustical universal structure," in *Proc. International Conference on Audio, Language and Image Processing (ICALIP)*, Jul. 2012, pp. 454-458.

[8] X. Chen, W. Q. Zhang, and J. Liu, "An improved model for voice conversion based on Gaussian mixture model," in *Proc. International Conference on Computer Application and System Modeling (ICCASM)*, 2010.

[9] D. Erro, I. Sainz, I. Saratxaga, E. Navas, and I. Hernaez, "MFCC+F0 extraction and waveform reconstruction using HNM: Preliminary results in an HMM-based synthesizer," in *Proc. FALA*, Vigo, 2010.

[10] P. R. Gurumoorthy, "LPC based voice morphing," University of Florida, 2006.

[11] A. Chadha, B. Savardekar, and A. Padhya, "Analysis of a modern voice morphing approach using gaussian mixture models for laryngectomees," *International Journal of Computer Applications*, vol. 49, no. 21, Jul. 2012

[12] Z.-H. Jian and Z. Yang, "Voice conversion using viterbi algorithm," in *Proc. International Symposium on Intelligent Signal Processing and Communication Systems*, Dec. 2007.

[13] J. Nirmal, P. Kachare, S. Patnaik, and M. Zaveri, "Cepstrum liftering based voice conversion using RBF and GMM," in *Proc. International Conference on Communication and Signal processing (ICCSP)*, 2013, pp. 570-575.

[14] C. Orphanidou, I. M. Moroz, and S. J. Roberts, "Multiscale voice morphing using radial basis function analysis," Oxford Center for Industrial and Applied Mathematics, University of Oxford, 2006.

[15] J. Nirmal, S. Patnaik, M. Zaveri, and P. Kachare, "Complex cepstrum based voice conversion using radial basis function," *ISRN Signal Processing*, vol. 2014, Feb. 2014.

Abdul Qavi (Islamabad, Pakistan, May 11, 1987) has recently done MS in Computer Engineering (2014) from Center for Advanced Studies in Engineering (CASE), Islamabad, Pakistan. He obtained his BS in Computer Science in 2009 from Kohat University of Science and Technology (KUST), Kohat, Pakistan where he was awarded Gold Medal. His area of specialization is DSP, speech/image processing, parallel processing, and machine learning.
He has worked with companies like And-Or Logic (www.andorlogic.com) and TEO International (www.teo-intl.com). He is currently working as Software Manager at TEO where he looks after various software projects.
Mr. Qavi is a member of Sciei and IEEE societies. He has been awarded two gold medals in his studies, and has been Dy. President of Computer Science Society at Kohat University of Science and Technology, Pakistan.

Kashif Basir has received an MS in Software Engineering from College of Electrical and Mechanical Engineering, National University of Science and Technology (CEME, NUST) in 2014. He obtained BS in Computer Science from Kohat University of Science and Technology, Pakistan. His research interests include component based software development and data mining in particular and artificial intelligence in general.
He has worked for Histone Solutions Islamabad as a Software Developer, where he worked on various software projects.

Shoab A. Khan did his PhD in Digital Signal Processing from Georgia Institute of Technology in 1995. He has worked with companies like Scientific Atlanta, Ingersoll Rand, and Cisco Systems in the area of signal processing and communication systems. He is a founder of a garage base startup Avaz Networks, now with $17 Million in venture funding and Center for Advanced Research in Engineering. The company has successfully developed a satellite burst modem for STM wireless and seven first time right ASICs for NEC, Scientific Atlanta, ITEX, and Nortel. Recently the company has delivered the industry highest density Media Processor for Voice over Packet market, a complex System on Chip with 12 processors on it. Dr. Khan is CTO and lead designer of this chip. He is the author of Digital Design of Signal Processing System: A network Approach published by John Wiley & Sons, Ltd, 978-0-470-74183-2, January 2011.