IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 23, 2021, accepted January 28, 2021, date of publication February 3, 2021, date of current version February 11, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3056711

Towards More Efficient DNN-Based Speech Enhancement Using Quantized Correlation Mask

SALINNA ABDULLAH, (Graduate Student Member, IEEE), MAJID ZAMANI, (Member, IEEE), AND ANDREAS DEMOSTHENOUS, (Fellow, IEEE)

Department of Electronic and Electrical Engineering, University College London (UCL), London WC1E 7JE, U.K.

Corresponding author: Salinna Abdullah (salinna.abdullah.13@ucl.ac.uk)

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/R512400/1.

ABSTRACT Many studies on deep learning-based speech enhancement (SE) utilizing the computational auditory scene analysis method typically employs the ideal binary mask or the ideal ratio mask to reconstruct the enhanced speech signal. However, many SE applications in real scenarios demand a desirable balance between denoising capability and computational cost. In this study, first, an improvement over the ideal ratio mask to attain more superior SE performance is proposed through introducing an efficient adaptive correlation-based factor for adjusting the ratio mask. The proposed method exploits the correlation coefficients among the noisy speech, noise and clean speech to effectively re-distribute the power ratio of the speech and noise during the ratio mask construction phase. Second, to make the supervised SE system more computationally-efficient, quantization techniques are considered to reduce the number of bits needed to represent floating numbers, leading to a more compact SE model. The proposed quantized correlation mask is utilized in conjunction with a 4-layer deep neural network (DNN-QCM) comprising dropout regulation, pre-training and noise-aware training to derive a robust and high-order mapping in enhancement, and to improve generalization capability in unseen conditions. Results show that the quantized correlation mask outperforms the conventional ratio mask representation and the other SE algorithms used for comparison. When compared to a DNN with ideal ratio mask as its learning targets, the DNN-QCM provided an improvement of approximately 6.5% in the short-time objective intelligibility score and 11.0% in the perceptual evaluation of speech quality score. The introduction of the quantization method can reduce the neural network weights to a 5-bit representation from a 32-bit, while effectively suppressing stationary and non-stationary noise. Timing analyses also show that with the techniques incorporated in the proposed DNN-QCM system to increase its compactness, the training and inference time can be reduced by 15.7% and 10.5%, respectively.

INDEX TERMS Correlation coefficients, deep neural network, dynamic noise-aware training, quantization, speech enhancement, training targets.

I. INTRODUCTION

Speech enhancement (SE) is the task of separating speech from nonspeech noise, used with the aim to improve perceived quality and intelligibility of speech. It is a procedure fundamental in signal processing for a wide range of applications including but not limited to hearing prostheses, such as hearing aids [1] and cochlear implants [2], mobile telecommunication [3], and robust automatic speech recognition [4]. Traditional SE methods such as spectral subtraction [5], Wiener filter [6], short-time spectral amplitude estimators [7], and maximum-likelihood spectral amplitude [8] algorithms have demonstrated good noise suppression performance when the assumed characteristics of the speech and noise signal are maintained. However, the performance of the traditional SE methods degrades considerably when presented with non-stationary noise or noisy speech at low signal-to-noise ratios (SNRs) since it is difficult to estimate the speech and noise properties in these conditions effectively.

The utilization of a mask for supervised SE was inspired by the concept of time-frequency (T-F) masking in the computational auditory scene analysis (CASA) method [9]. The CASA method operates based on the auditory perception

The associate editor coordinating the review of this manuscript and approving it for publication was Shuping He.

mechanism, exploiting grouping cues such as pitch and onset without assuming any properties or models of the noise. T-F masking involves applying a two-dimensional weighting to the T-F representation of a noisy speech to separate the clean speech. The most common T-F mask is the ideal binary mask (IBM) [10], which denotes whether a T-F unit is dominated by the target signal by taking up values of one and zero. The binary decision is usually computed by comparing the local SNR in each T-F unit against a predefined threshold. From listening studies conducted, the IBM showed improved speech intelligibility for both normal-hearing and hearing-impaired listeners in noisy conditions [11], [12]. An improvement from the IBM is the ideal ratio mask (IRM), which can be viewed as a soft version of the IBM as it adopts continuous values between zero to one to represent the probability of a T-F unit being target dominant instead of hard labels of strictly ones or zeros [13], [14]. The IRM has been shown to improve objective speech quality in addition to predicted speech intelligibility over the IBM as the latter method has a drawback of often wrongly removing speech or retaining noise portions, introducing speech distortion or residual musical noise [15].

One of the recently proposed training targets is the optimal ratio mask (ORM) [16]. The ORM, which considers the spectral coherence between the speech and noisy in a noisy speech mixture, has been reported to have the potential to improve the SNR by approximately 3 dB over the IRM by theoretical analysis, implying a better separation performance. Besides the ORM, training targets that incorporate phase information such as the complex ideal ratio mask [17] and phase sensitive mask [18] have also been utilized in SE recently. The complex ideal ratio mask is computed on the complex domain whereas the phase sensitive mask introduces phase information and operates on the real domain. Favourable results have been reported for these training targets, but the inclusion of phase information increases the difficulty in estimating the clean speech.

Over the last decade, deep learning has been widely utilized for SE and has demonstrated exceptional denoising capability even in challenging conditions such as when dealing with non-stationary noise, unseen and low SNR conditions. Supervised deep neural networks (DNNs), which include feedforward multilayer perceptrons [15], convolutional neural networks [19], recurrent neural networks [20] and generative adversarial networks [21], in particular, have significantly elevated the performance of SE through capturing the complicated relationship between the noisy speech and clean speech. There has also been an increase in the use of modified versions of conventional neural networks. For example, Takeuchi *et al.* [22] proposed a real-time SE system using equilibrated recurrent neural network to solve the problem of vanishing or exploding gradient without increasing the number of parameters within the network.

Although deep learning-based frameworks have outstanding capability to be exploited in training ratio masks, their large storage space requirements render them unsuitable to implement in devices with limited resources. To improve the compactness of SE models, methods such as pruning [23], sparse constraints [24] and quantization [25] have been proposed. Wu *et al.* report 90.24% size reduction in their SE model from their baseline with their proposed combinative approach that encompasses parameter pruning, parameter quantization and feature-map quantization techniques [26]. Ko *et al.* investigated the correlation of precision scaling and neuron numbers in an SE model and found that neural networks with lower bit precision significantly reduce the processing time by up to 30x. However, their performance impact is significantly deteriorated ($<$3.14%) when implemented in classification tasks such as those present in voice activity detection [27]. In [28], a two-stage quantization approach was derived to reduce the number of bits required to represent floating-point parameters. Dynamic quantization has been recently explored by Chang and Liu [29] and Xiong *et al.* [30] for vehicle systems, which suggests a growing trend in the use of an adaptive approach for signal quantization. In [29], a dynamic quantizer was developed to adjust the quantization level and parameter used to reduce the steady state limit cycle in in-vehicle networked systems as static quantizers suffer from sensor failure and dropouts. In [30], the control input of nonfragile feedback control for active suspension in vehicles is quantized by a class of quantizers with adjustable dynamic parameters to provide more desirable closed-loop system performance.

This paper explores a new ratio mask representation for the supervised SE to attain improved denoising capability and implements efficiency optimization techniques to make the SE algorithm more appropriate for real-time applications on compact devices such as cochlear implants. The main contributions of the proposed SE method include the following: 1) it utilizes a correlation-based learning target where the IRM is more optimally adjusted with inter-channel correlation (ICC) factors. This paper proposes a more efficient method of calculating normalized cross-correlations (NCCs) in the ICC factors through using sum tables. The computational cost (or time) of the proposed sum tables method is significantly lower than direct calculation using the ICC definition and this is achieved without sacrificing the SE performance as the energy proportions between speech and noise are used to more precisely adjust the ratio mask; 2) fixed and k-means quantization techniques are used in a two-stage process to reduce the required number of bits to represent the neural network weights, learning target and acoustic features in order to make the supervised deep learning-based SE more compact; and finally, 3) a DNN network has been optimized through experimental findings to be used in combination with the proposed quantized correlation mask (QCM). In this study, the SE capability of the proposed QCM is evaluated when implemented in conjunction with a four-layer structured DNN (DNN-QCM). The incorporation of dropout regulation, Boltzmann pre-training and noise-aware deep mapping strategies in this work introduced a more robust SE with

improved generalization performance on various scenarios. Through a series of experiments using datasets combining the TIMIT [31] corpus and the NOISEX-92 [32] database, the DNN-QCM provided better perceived speech intelligibility and quality scores when compared to a standard DNN-based SE using IRM (DNN-IRM) and another correlation-based training target proposed in [33] (DNN-CRM, with CRM being the acronym for correlation ratio mask), as well as a minimum mean square error estimator-based (MMSE-based) and non-negative matrix factorization-based (NMF-based) SE.

The rest of this paper is organized as follows. In Section II, the principles of the proposed method for calculating low cost ICC factors are explained, and the quantization techniques and acoustic feature extraction methods are presented. Section III elaborates the DNN framework and setup employed for the SE application. The experimental settings, including the datasets, evaluation metrics and benchmark models, are presented in Section IV. The experimental results obtained and details on the complexity, convergence and timing analyses of the DNN-based methods are also provided and discussed in this section. Concluding remarks are drawn in Section V.

II. THE PROPOSED SYSTEM

The block diagram of the proposed DNN-QCM is shown in Fig. 1. The baseline system is constructed in two stages. In the training stage, a feedforward DNN with four hidden layers is trained with acoustic features (feature extraction is outlined in Section II-D), which is a combination of the amplitude modulation spectrogram (AMS), relative spectral transformed perceptual linear prediction (RASTA-PLP) coefficients, mel-frequency cepstral coefficients (MFCCs) and 64-channel gammatone frequency cepstral coefficients (GFCCs). These acoustic features are extracted from the noisy speech mixture. For the generation of the training target, the ICC factors are calculated from the gammatone magnitude spectra among the clean speech, noise and noisy speech mixture, and then combined with the channel-weight contour extracted from the noisy speech mixture. The inclusion of the ICC factors fine-tunes the IRM with information about the correlation between the noisy speech with noise and noisy speech with clean speech. This provides the DNN with correlation-based training on top the power spectra-based training given by the IRM. Dynamic noise-aware training (DNAT) is achieved by estimating the NCC between the noisy speech and noise in each gammatone channel on a frame-by-frame basis. The NCC coefficients are then fed together with the training target and acoustic features to the DNN for training. The DNAT elevates the DNN's SE performance and allows it to provide better denoising capability when dealing with non-stationary and unseen noises. Quantization represented by 'Q' blocks in Fig. 1 is applied to the weight parameters within the DNN and to the acoustic features and training target to reduce the size of the SE model with more efficient data representation. The enhancement stage then

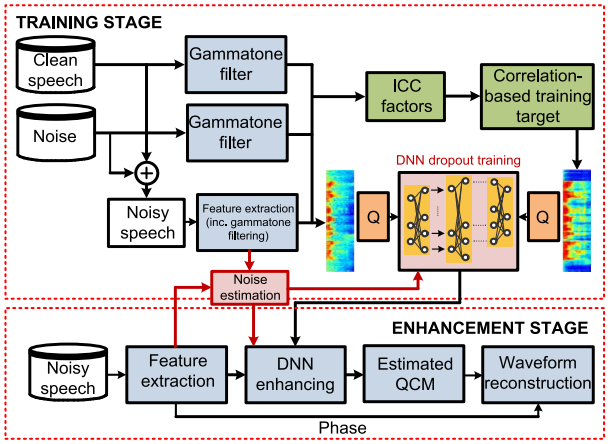


FIGURE 1. Block diagram of the proposed DNN-QCM, where the main novelties including the correlation-based training target, correlation-based DNAT (the noise estimation block) and quantization processes are presented in blocks shaded in green, red and orange, respectively. 'Q' denotes the quantization techniques which are applied to make the DNN model more compact.

involves feeding the acoustic features of the noisy speech to the trained DNN model to estimate the ratio mask for reconstructing the clean speech signal (i.e., the enhanced signal). The proposed DNN-QCM embeds both improved denoising performance as well as compactness to the SE framework which are crucial in real-time applications.

A. CORRELATION-BASED TRAINING TARGET

For improved readability, the notations used in this section are listed and defined in Table 1. In this study, the IRM is effectively extended to consider the ICC among the noise, clean speech and noisy speech signal. The IRM can be viewed as a soft version of the IBM. Both masks are defined on a two-dimensional T-F representation of a noisy speech signal such as a cochleagram or spectrogram. However, instead of assigning a '1' if the SNR within a T-F unit exceeds the specified threshold and '0' otherwise, the IRM adopts a more partial suppression of T-F units that are deemed noisy. The optimal estimator of the power spectra for speech and noise is similar to the square-root Wiener filter, which is defined as when the tuneable parameter (β), used to adjust the scale of the mask, is set to 0.5 in the IRM definition [15]. With the ICC factors introduced, the IRM definition becomes [33]:

$$ICC_{IRM\,(c,m)} = \frac{\rho_x(c,m)\cdot P_x(c,m)}{\rho_x(c,m)\cdot P_x\,(c,m) + \rho_n(c,m)\cdot P_n(c,m)}, \quad (1)$$

where $P_x(c,m)$ and $P_n(c,m)$ denote the speech and noise energy, respectively, of the mth frame in the cth channel. $\rho_x(c,m)$ is the NCC coefficient between the clean speech and noisy speech power spectra in the cth channel of the mth frame and $\rho_n(c,m)$ is the NCC coefficient between the noise and noisy speech power spectra in the cth channel of the mth frame. The ICC factors are expressed as percentages of the clean speech or noise components within the noisy speech signal. When speech components are dominant in a channel

TABLE 1. Index of key notations.

Notation	Description
c	Gammatone channel index
m	Frame index
u	Window origin
W	Window size
l	Number of samples in a signal
τ	Shift between windows
T	Matrix transpose
$y_{c,m}$	Magnitude spectrum of noisy speech
$x_{c,m}$	Magnitude spectrum of clean speech
$n_{c,m}$	Magnitude spectrum of noise
β	Tuneable parameter for IRM
$ICC_{IRM(c,m)}$	New IRM definition with ICC introduced
$P_x(c,m)$	Energy of clean speech
$P_n(c,m)$	Energy of noise
$\rho_x(c,m)$	NCC coefficient between the clean and noisy speech
$\rho_n(c,m)$	NCC coefficient between the noise and noisy speech
$s_{y,x}(m,\tau)$	Sum tables of cross-correlation $y_{c,m}^T \cdot x_{c,m}$
$s_{y,n}(m,\tau)$	Sum tables of cross-correlation $y_{c,m}^T \cdot n_{c,m}$
$s_y^2(\mathbf{m})$	Sum tables of $\|y_{c,m}\|^2$
$s_x^2(\mathbf{m})$	Sum tables of $\|x_{c,m}\|^2$
$s_n^2(\mathbf{m})$	Sum tables of $\|n_{c,m}\|^2$

of a frame, $\rho_x(c, m)$ becomes larger as stronger correlation exists between the speech components and the noisy speech. Similarly, $\rho_n(c, m)$ is larger when noise components are dominant. Therefore, with the adaptive ICC factors, the weights of speech and noise components can be adjusted according to how much they are present in a noisy speech unit.

Reference [33] recommends calculating in the adaptive ICC factors, $\rho_n(c, m)$ and $\rho_x(c, m)$, by means of NCCs in the gammatone domain (i.e., with 64 channels, used to process the audio signals sampled at 8 kHz). The ICC factors can then be expressed as:

$$\begin{cases} \rho_n(c,m) = \dfrac{y_{c,m}^T \cdot n_{c,m}}{\sqrt{\|y_{c,m}\|^2 \cdot \|n_{c,m}\|^2}}, & (a) \\ \rho_x(c,m) = \dfrac{y_{c,m}^T \cdot x_{c,m}}{\sqrt{\|y_{c,m}\|^2 \cdot \|x_{c,m}\|^2}}, & (b) \end{cases} \tag{2}$$

where $y_{c,m}$, $n_{c,m}$ and $x_{c,m}$ are the magnitude spectrum column vectors of noisy speech, pure noise and clean speech in each frame of each gammatone channel, respectively. T represents the transpose of the matrix.

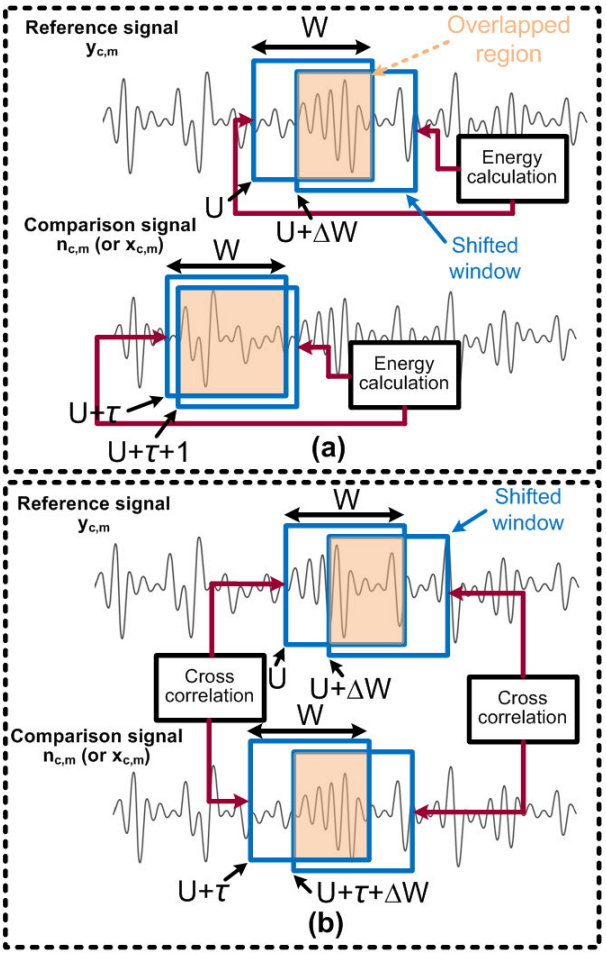


FIGURE 2. A diagram depicting the NCC calculation process where the energy calculation of the reference and comparison signals are shown in (a), and the non-normalized cross-correlation calculation between the reference and the comparison signals are shown in (b). In (a), the extensive search of the comparison windows (e.g., when the search is equal to τ and $\tau + 1$) and the shifting of the reference window at less than W lead to significant overlap (indicated by the shaded overlapping regions between the shifted windows) where redundant calculations are present. Similarly in (b), overlap persists as long as the comparison signal is shifted by less than W, allowing the non-normalized cross-correlation to be calculated more efficiently by eliminating redundant calculations.

The noisy speech $(y_{c,m})$ from (2) is also referred to as the reference signal and the $n_{c,m}$ (or $x_{c,m}$) is referred to as the comparison signal. The process of calculating the correlation between two signals is depicted in Fig. 2. Fig. 2(a) shows a segment of a sample speech signal where the reference window is located within the interval of $[u, u + W - 1]$, where u is the origin of the reference window and W denotes the window size. As interpreted from (2-a), the NCC is obtained by calculating the energy (i.e. squared signal) of the reference signal $\sum_{l=u}^{u+W-1} y_{c,m}(l)^2$ and the comparison signal $\sum_{l=u}^{u+W-1} n_{c,m}(l+\tau)^2$ in the denominator as well as the dot product between the signals in the numerator. Here m, denoting the frame number, has been expanded and replaced with l to account for overlap (i.e., l is the number of samples in a signal). τ is the shift between the comparison and reference windows. When calculating the NCCs, the reference window is shifted by ΔW ($\Delta W \ll W$) through the

entire speech signal, and for each reference window, (2) is calculated progressively for every search τ. Such calculation process requires significant computational cost and is time-consuming.

In place of performing the costly NCC calculations in (2), this paper proposes performing NCCs using pre-calculated sum tables [34] to reduce the operations required to obtain the ICC factors for adjusting the ratio mask. As shown in Fig. 2(a), the sum tables-based NCC exploits the fact that most calculations are redundant due to the extensive search of the comparison windows (i.e., $n_{c,m}$ and $x_{c,m}$) and high overlap between the reference windows (i.e., $y_{c,m}$). When the searches for the comparison windows are equal to τ and $\tau + 1$, significant overlap such as the one depicted by the shaded area in Fig. 2(a) occurs. This is seen as the foundation of the sum tables method: redundant operations in the overlapped region can be eliminated with more efficient energy subtraction between the starting and ending points of the window.

The sum tables for the comparison window $s_n^2(m)$ (or $s_x^2(m)$) are constructed as follows:

$$s_n^2(m) = \begin{cases} n^2(m) + s_n^2(m-1) & (1 \le m \le M) \\ 0 & (m = 0), \end{cases} \quad (3)$$

where m, again, represents the frame number and M is the total length in frames of the comparison signal. The energy of the comparison window used in the denominator of (2-a) can then be calculated as:

$$\sum_{l=u}^{u+W-1} n^2(l+\tau) = s_n^2(u+W-1+\tau) - s_n^2(u-1+\tau), \quad (4)$$

where u is the origin of the window, W is the window length and τ is the shift between the comparison and reference windows.

Similarly for the reference windows, energy differencing using the sum tables constructed for the reference signal can be used to calculate the energy of the reference signal $\sum_{l=u}^{u+W-1} y_{c,m}(l)^2$. Hence, the sum tables for $s_y^2(m)$ and the energy of the reference window $y^2(l)$ can be constructed similarly to (3) and (4), respectively.

Finally, the same theory of overlap and redundancy is employed to define the standard (i.e., non-normalized) cross-correlation terms in the numerator of (2-a) and (2-b). The numerators in (2-a) and (2-b) are the products of the reference and the comparison signals. As shown in Fig. 2(b), during the whole shift and product process between the two signals, there is a remarkable overlap between the reference window at u and the comparison window at $u + \tau$, and the shifted reference window at $u + \Delta W$ and the comparison window at $u + \tau + \Delta W$. This leads to the following set of sum tables:

$$s_{y,n}(m,\tau) = \begin{cases} y(m)\cdot n(m+\tau) + s_{y,n}(m-1,\tau) & (1 \le m \le M) \\ 0 & (m = 0). \end{cases} \quad (5)$$

The numerator can then be calculated through the subtraction of the sum tables $s_{y,n}(m,\tau)$ before and after shifting:

$$\sum_{l=u}^{u+W-1} y(l)n(l+\tau) = s_{y,n}(u+W-1,\tau) - s_{y,n}(u-1,\tau). \quad (6)$$

Therefore, the ICC factor $\rho_n(m)$ in (2-a) can be calculated through sum tables $s_n^2(m)$ and $s_y^2(m)$ for the denominator, and $s_{y,n}(m,\tau)$ for the numerator (i.e. $\rho_n(m) \approx s_{y,n}(m,\tau) / \sqrt{s_n^2(m) . s_y^2(m)}$). Similar applies to the ICC factor $\rho_x(m)$ in (2-b).

B. QUANTIZATION TECHNIQUES

In this section a two-stage pipeline quantization approach as shown in Fig. 3 is proposed to optimally reduce the number of bits required to represent parameters encoded in floating point representations. The first stage utilizes fixed quantization to retain the necessary number of bits to represent the training target and acoustic feature set including the noise estimate without compromising the SE performance. On the other hand, the second stage comprises a k-means-based quantization approach where the weights of the neural network are replaced with quantized centroids.

In the first stage where the fixed quantization is concerned, the impact on the evaluation metrics comprising the short-time objective intelligibility score (STOI) [35], perceptual evaluation of speech quality (PESQ) [36], segmental SNR (SSNR, in dB) [37] and the log-spectral distortion (LSD, in dB) [38] are observed when the number of fixed quantization levels is increased incrementally by 1 for levels 1 to 10 and then incrementally by 5 for levels 10 to 50. All STOI, PESQ and LSD but the SSNR results shown in Fig. 4 suggest that quantization levels of 20 and above are suitable. This means that the 14-bit representation used previously can be reduced to a 5-bit representation, which gives up to 32 quantization levels corresponding to the quantized mask values. The SSNR may be a poor indicator of the SE performance as the quantization may have led to more stringent noise suppression that also came with high distortion. A similar process was applied to the acoustic feature set and it was found that a 5-bit representation provides sufficient granularity in the feature set without causing significant deterioration in the SE outcome. However, an additional sign bit is required for the representation of the acoustic feature set.

Utilizing the k-means algorithm in the second stage, the weights in a neural network are grouped into several clusters, with each cluster of weights sharing a centroid value. With the k-means algorithm, meaningful centroids often representing the more dominant weight values are extracted and used in place of any numbers in the 32-bit float range. An example of the k-means based parameter quantization process is depicted in Fig. 5. With k = 4, a look-up table with 4 cluster centroids is obtained. Each weight in the model is then denoted with a cluster index that corresponds to its cluster. Therefore, with a look-up table containing 4 integer indices representing the

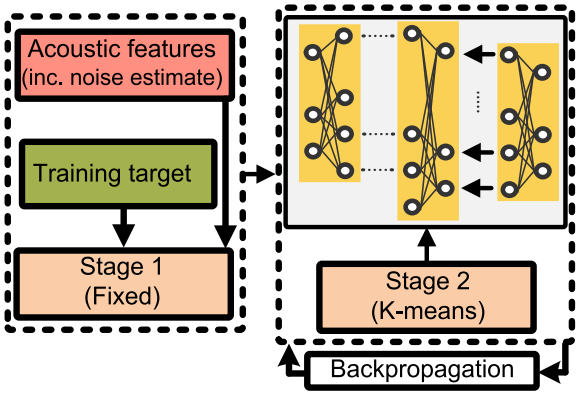


FIGURE 3. Block diagram depicting the two-stage quantization process where a fixed quantization is applied on the acoustic features and training target in Stage 1, and a k-means quantization is then applied on to the neural network weights in Stage 2.

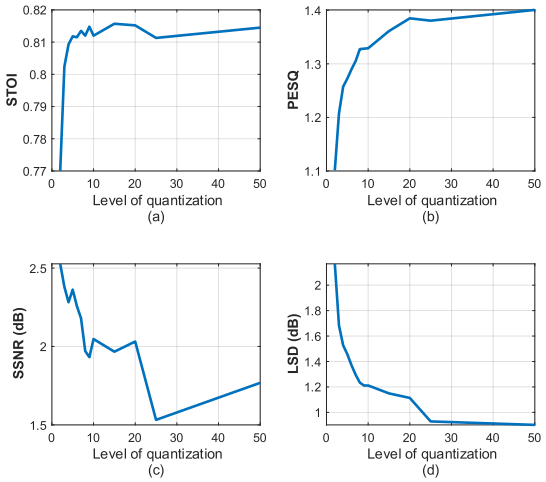


FIGURE 4. The resulting average performance evaluation when different levels of fixed quantization were applied to the QCM.

4 centroids, only 2 bits are required to represent the weights in the neural network model. For finding the optimal 'k' value, the number of clusters was set to 2, 4, 8, 16, 32 and 64, and the corresponding STOI and PESQ results were obtained as shown in Fig. 6. It was found that k-means with k $=$ 32 ensures minimal bit requirement without compromising the performance of the SE in this work. Therefore, the 32-bit floating number representation can also be replaced with a 5-bit integer representation, increasing the compactness of the DNN model.

C. ACOUSTIC FEATURES

The acoustic features used in this study are a fixed set of complementary features recommended in [39]. The complementary feature set is a concatenation of the: (1) AMS; (2) RASTA-PLP coefficients; (3) MFCCs and; (4) GFCCs. In the computation of the AMS feature, the envelope of the signal is extracted using full-wave rectification before

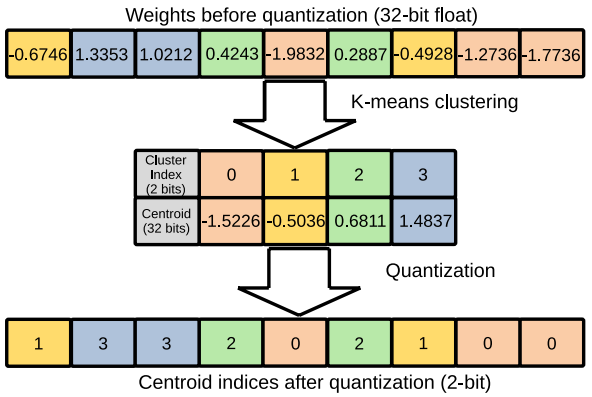


FIGURE 5. An example of the weight quantization technique where 4 centroids were obtained utilizing the k-means algorithm.

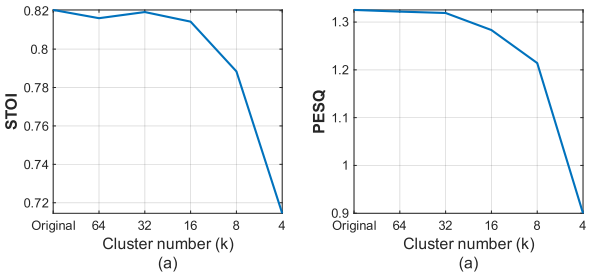


FIGURE 6. The resulting average performance evaluation when different values of 'k', representing the number of centroid clusters, were used to quantize the weights of the DNN.

decimation by a factor of 4 is applied. Thereafter, the decimated envelope undergoes Hanning windowing before being zero-padded to become a 256-point FFT. Finally, the resulting FFT magnitudes are integrated by 15 triangular windows uniformly spaced from 15.6 to 400 Hz to produce a 15-D AMS feature vector. To derive the RASTA-PLP coefficients, the power spectrum of the signal is warped to a 20-channel Bark scale using trapezoidal filters, log-compressed, filtered by the RASTA filter [40] with a single pole at 0.94 and then expanded again by an exponential function. Subsequently, loudness pre-emphasis and intensity loudness law are applied. The cepstral coefficients from linear predictions then form the RASTA-PLP features. Following a common practice in speech recognition, a 12th order linear prediction model is used. This yields a 13-D (including the zeroth cepstral coefficient) RASTA-PLP feature vector. For the MFCCs, the signal is first pre-emphasized, followed by a 512-point short-time Fourier transform with a 20-ms Hamming window. The resulting power spectra are then warped to a 64-channel mel scale. This is then followed by a log operation and discrete cosine transform (DCT) to yield a 31-D MFCC feature vector. The extraction of the GFCC features involves decomposing the signal with a 64-channel gammatone filterbank first before being decimated to an effective sampling rate of 100 Hz. The outputs are then loudness-compressed by a cubic root

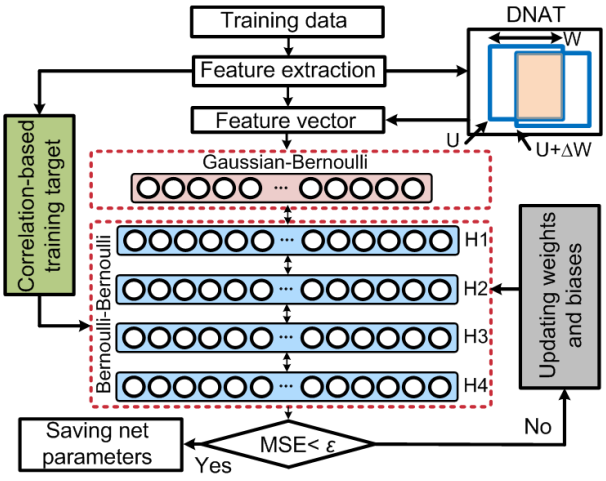


FIGURE 7. Illustration of the DNN training framework. The neural networks shown are the RBM stack used for pre-training which includes a visible Gaussian-Bernoulli RBM layer with four hidden Bernoulli-Bernoulli RBM layers (i.e. H1, H2, H3 and H4). The RBM stack can be replaced with feedforward multilayer perceptrons to depict the subsequent main training process which occurs after the network weights and biases have been initialized with the RBM.

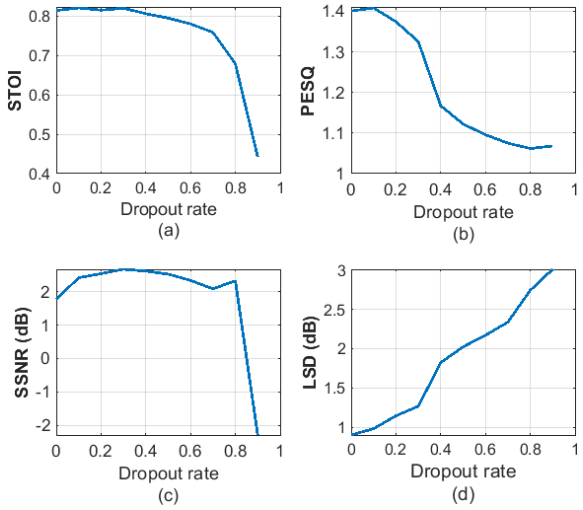


FIGURE 8. The impact of dropout regulation on SE performance.

operation, followed by DCT to yield a 31-D GFCC feature vector, as with the MFCC feature vector.

III. DNN-BASED SPEECH ENHANCEMENT

The acoustic features extracted from a noisy speech mixture along with the corresponding training targets are fed into the DNN for training as shown by the flowchart in Fig. 7. An experimental study was conducted to optimize the configuration of the feedforward DNN within the proposed SE system for the mapping process. In the preliminary study, a sample subset of the training and testing dataset (further described in Section IV-A) were used to evaluate the SE performance when changes to the configuration of the DNN were made. As shown in the sample results, it was observed that the utilization of dropout regulation (see Fig. 8) for the hidden layers led to improvements across all evaluation metrics (explained in Section IV-B). It was found through the simulations that a dropout rate with 0.2 provides the optimal SE performance.

The final DNN configuration consists of four hidden layers, each having 1024 Rectified Linear Units (ReLUs) as their activation functions. The ReLU activation function is responsible for transforming the summed weighted input from a node into a non-linear output by being linear for all positive values and returning 0 for all negative values. Mathematically, the ReLU transformation function is given by:

$$y = \max(0, x)\,. \tag{7}$$

At the output layer, a sigmoid activation function is used since the training targets possess a range between [0,1].

Layer-wise pre-training using restricted Boltzmann machines (RBMs) [41] is employed to make the feature learning more robust by more optimally initialize the network's weights and biases with a subset of the training data. This paper utilizes a multiple of RBMs stacked together for pre-training as illustrated in Fig. 7. The first RBM is a Gaussian-Bernoulli RBM that has one visible layer of linear variables connected to a hidden layer. Thereafter, a pile of hidden Bernoulli-Bernoulli RBMs (i.e. H_1, H_2, H_3 and H_4) is added to the prior Gaussian-Bernoulli RBM. The RBMs are trained by contrastive divergence algorithm [42] for faster and more computationally stable convergence. 100 epochs of mini-batch gradient descent are employed for RBM pre-training and 100 epochs of limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm [43] are used for fine-tuning the whole network. Lastly, a learning rate of 0.001 is used for the Gaussian-Bernoulli RBM and a learning rate of 0.01 is used for the Bernoulli-Bernoulli RBM.

To improve noise robustness of the SE system through introducing DNAT, noise information is appended to the input acoustic feature and fed to the DNN to better the prediction of the clean speech as shown in Fig. 7. Static noise-aware training used in [44] that assumes a fixed noise estimation over the entire utterance, is not a practical approach for suppressing non-stationary or burst noises. For the suppression of non-stationary or burst noises, dynamic noise estimation is needed and this can be implemented using the conventional MMSE-based noise estimation method [45] at each frame. However, it has a drawback of introducing non-linear distortion in the estimated noise spectrum which in turn results in more challenging DNN learning. Therefore, DNAT employing the correlation-based approach discussed in Section (II-A) is proposed here by the use of the following equation:

$$r_{yn}(m) = r_{yy}(m) - r_{yx}(m)), \tag{8}$$

where $r_{yy}(m)$ is the autocorrelation of the noisy speech $y(c, m)$ and $r_{yx}(m)$ is the cross-correlation between the noisy speech and the estimated clean speech. $r_{yn}(m)$, which is

the cross-correlation between the noisy speech and noise, is calculated for each channel of the gammatone frequency domain and then concatenated to the acoustic feature vector for DNN training to reflect the coherent relationship among different frames and gammatone frequency channels.

After initializing the network's weights and biases, the DNN is trained by streaming the feature vectors through a standard backpropagation algorithm where the mean squared error (MSE) is used as the cost (loss) function:

$$E\left(w^i, b^i\right) = \frac{1}{2}\sum_m \sum_c \left\| \left[Q\hat{C}M(c,m), w^i, b^i \right] - a^i(c,m) \right\|^2, \tag{9}$$

where c is the channel index, m is the frame index and i represents the network layer index. $Q\hat{C}M\ (c,m)$ denotes the proposed quantized training target in this paper. $a^i\ (c,m)$ is the activation value in the cth channel of the mth frame. Finally, w^i and b^i are the weights and biases vectors in the ith layer, respectively. The update and estimation of w^i (and similarly, b^i) in the ith layer, with a learning rate $\lambda = 0.1$, can then be completed iteratively according to Eq. (10) along with a momentum factor ω:

$$\begin{cases} w^i \leftarrow w^i + \Delta w_t^i \\ \Delta w_t^i = -\lambda . E\nabla\left(w^i\right) + \omega \cdot \Delta w_{t-1}^i, \end{cases} \tag{10}$$

where t is the index of the iteration.

Adaptive gradient descent [46] along with a momentum term (ω) are used as the optimization technique. In this paper, a momentum rate of 0.5 is initially used for the first 5 epochs before it is increased to 0.9 thereafter. The total number of epochs is 100. During each iteration, the difference between the training target and the concatenated input feature vector is fed back through the network to generate the mapping patterns in the hidden and output layers.

IV. RESULTS AND ANALYSIS

A. DATASETS

In the performance evaluation, 1500 randomly chosen clean utterances from the TIMIT [31] training set were used as the training utterances and 100 utterances from the TIMIT core test set, which comprises of 192 utterances from unseen speakers of both genders, were used as the test utterances. Five types of noises including babble, factory, pink, Volvo (car) and white noise from the NOISEX-92 [32] database were used as the training noises. The same five noises were used for testing, with the addition of 'f16' and 'factory 2' from the same database used for evaluating the generalization performance. To avoid using the exact same frames of noise for both testing and training, random cuts of the first 2 minutes of each noise were used for training whereas random cuts of the last 2 minutes of each noise were used for testing. A sampling frequency of 8 kHz was used throughout the experiment. The noisy speech mixtures were generated by contaminating the utterances with each type of noise at -5, 0 and 5 dB SNR.

B. EVALUATION METRICS

Since the output from the speech enhancement system is ultimately presented to a human listener, the evaluation metrics used have been designed to quantitatively predict how a human listener will perceive the enhanced speech signal. Speech intelligibility and quality are two different key aspects of speech perception that are often assessed both objectively and subjectively to evaluate the performance of an SE system. The STOI [35] is the most commonly used intelligibility metric since it has been shown to be highly correlated with human speech intelligibility score consistently. The STOI measures the correlation between the short-time envelopes of the clean signal and the enhanced signal, and it is presented in a range between 0 and 1, which can be interpreted as percentage correct. For speech quality, the PESQ [36] is recommended by the International Telecommunication Union as a standard metric. As with most speech intelligibility and quality measures developed, the PESQ is also calculated by comparing the enhanced speech with the clean reference speech. More specifically, PESQ compares the loudness spectrum of the two signals to produce a score in a range of -0.5 to 4.5. Other evaluation metrics used are the SSNR [37], representing the degree of noise reduction, and the LSD [38], which denotes the level of speech distortion present in the enhanced signal. For the LSD, lower values equate to lower levels of speech distortion, thus indicating better SE performance.

C. EVALUATION OF DIFFERENT SE NETWORKS BASED ON GENERATED NOISY SPEECH SIGNALS

To assess the benefits of the proposed DNN-QCM, its SE performance was compared with: (1) a model-based SE using the MMSE by Hendriks *et al.* [47]; (2) a supervised NMF method in [48], where informative prior distributions obtained from the temporal dependencies of speech and noise signals are applied in a Bayesian framework to perform the NMF; (3) the DNN-IRM (when the conventional IRM is used as the training target); and (4) the DNN-CRM proposed in [33] where a conventional correlation-based training target is used for SE (DNN-CRM, with CRM being the acronym for correlation ratio mask). In the MMSE system, an MMSE estimator of speech DFT coefficients assuming a generalized gamma distribution for speech magnitude is employed.

Table 2 shows how the DNN-QCM compare to the MMSE, NMF, DNN-IRM and DNN-CRM-based SE. The SE algorithms were subjected to trained noise types (i.e., babble, factory, pink, Volvo and white noise) under -5, 0 and 5 dB SNR conditions. The values printed in boldface indicate the best achieved metric values for each noise type and SNR condition. The introduction of SE is vital for achieving improved speech intelligibility and quality. This is reflected in the improvements across all evaluation metrics over the unprocessed data after the SE systems were introduced. Among all the SE algorithms, the MMSE-based SE gave the worst performance as it consistently led to the

TABLE 2. Evaluation performance comparisons of the various SE algorithms, including the proposed DNN-based SE with QCM, on different noise types and SNR conditions.

SE	Babble			Factory			Pink			Volvo			White		
	-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB
	STOI														
Unprocessed	0.56	0.67	0.73	0.57	0.70	0.82	0.58	0.68	0.84	0.84	0.96	0.96	0.58	0.69	0.82
MMSE	0.58	0.69	0.78	0.58	0.68	0.90	0.60	0.71	0.84	0.86	0.95	0.96	0.62	0.70	0.83
NMF	0.61	0.74	0.81	0.65	0.73	0.93	0.71	0.76	0.93	0.86	0.97	**0.98**	0.70	0.73	0.86
DNN-IRM	0.77	0.84	0.87	0.80	0.81	0.96	0.83	0.85	0.95	**0.93**	0.97	**0.98**	0.79	0.84	0.93
DNN-CRM	0.84	0.91	0.90	0.86	0.91	**0.97**	0.86	0.92	0.95	**0.93**	**0.98**	**0.98**	0.83	0.92	0.95
DNN-QCM	**0.88**	**0.93**	**0.93**	**0.89**	**0.94**	**0.97**	**0.88**	**0.93**	**0.97**	**0.93**	**0.98**	**0.98**	**0.85**	**0.95**	**0.96**
	PESQ														
Unprocessed	1.07	1.09	1.45	1.03	1.06	1.58	1.05	1.08	1.61	1.35	1.53	1.57	1.06	1.07	1.52
MMSE	1.09	1.13	1.47	1.04	1.07	1.60	1.09	1.17	1.63	1.51	1.67	1.68	1.08	1.09	1.53
NMF	1.13	1.17	1.53	1.07	1.09	1.67	1.19	1.30	1.70	1.85	2.03	2.05	1.11	1.26	1.76
DNN-IRM	1.15	1.36	1.89	1.11	1.17	1.84	1.25	1.34	1.85	1.87	1.98	2.13	1.17	1.32	1.85
DNN-CRM	1.29	1.64	1.94	1.38	1.57	1.88	1.30	1.54	1.85	1.96	2.04	2.16	1.12	1.52	1.84
DNN-QCM	**1.34**	**1.79**	**2.00**	**1.43**	**1.66**	**1.89**	**1.33**	**1.55**	**1.88**	**2.01**	**2.14**	**2.19**	**1.18**	**1.58**	**1.88**
	SSNR (dB)														
Unprocessed	-6.07	-4.74	-2.08	-5.28	-4.60	-3.15	-5.14	-4.53	-3.21	-4.78	-4.11	-2.98	-5.11	-4.71	-3.43
MMSE	-3.50	-3.01	-1.50	-3.32	-1.98	-1.87	-3.22	-0.80	-0.43	-0.66	0.60	1.23	-1.33	-0.94	0.01
NMF	-2.50	-1.06	-0.56	-2.43	-0.22	0.02	-2.32	1.37	1.44	0.62	4.94	5.31	1.06	1.73	1.93
DNN-IRM	-0.87	2.12	3.43	-0.89	0.06	1.34	-0.84	1.54	1.86	4.58	6.54	6.89	1.32	1.93	2.61
DNN-CRM	**0.61**	**2.32**	3.99	0.92	1.51	1.64	0.92	2.18	2.47	**4.86**	**6.93**	6.90	1.89	2.51	2.74
DNN-QCM	**0.61**	2.22	**4.01**	**0.96**	**1.58**	**1.79**	**0.98**	**2.24**	**2.55**	**4.86**	6.89	**6.93**	**1.94**	**2.57**	**2.77**
	LSD (dB)														
Unprocessed	1.72	1.61	0.98	2.98	2.62	2.51	2.82	1.94	1.65	0.89	0.83	0.79	2.56	2.28	1.93
MMSE	1.61	1.48	0.92	2.60	2.26	2.12	2.55	1.62	1.53	0.90	0.95	0.75	2.33	1.76	1.67
NMF	1.78	1.45	0.84	2.14	2.02	1.89	2.09	1.49	1.45	0.85	0.81	0.65	1.87	1.52	1.55
DNN-IRM	1.02	0.97	0.67	1.76	1.47	1.13	1.73	0.99	0.89	0.74	0.56	0.52	1.34	1.10	1.01
DNN-CRM	0.91	0.86	**0.62**	1.00	0.97	0.89	1.22	0.91	0.85	**0.71**	0.54	**0.51**	1.11	0.99	0.91
DNN-QCM	**0.86**	**0.80**	**0.62**	**1.06**	**0.95**	**0.82**	**1.00**	**0.88**	**0.83**	**0.71**	**0.53**	**0.51**	**1.07**	**0.93**	**0.84**

lowest speech intelligibility and quality improvements in every scenario. This is because the model-based SE relies on its assumption on speech characteristics for denoising. In many cases, proposed DNN-QCM gave better evaluation metric scores across all noise types and SNR conditions. When processing speech contaminated with babble and Volvo noise at 0 dB, however, the DNN-CRM provided the best SSNR. The performance of the DNN-CRM often fell slightly short of the DNN-QCM. This suggests that although the sum tables method and quantization techniques were used to increase the efficiency and compactness of the DNN-QCM, the introduction of DNAT and nuances as a result of optimizing the DNN configuration have led to its net better performance. When denoising speech utterances contaminated with Volvo noise, the STOI scores obtained from all SE algorithms were almost equally high especially at 0 and 5 dB. The evaluation metric values obtained from processing the Volvo noise were higher than the rest of the noise because it is the more stationary noise compared to the other noise types. A one-way analysis of variance (ANOVA) performed to assess significant differences between the STOI scores obtained with each SE algorithm in Table 2 confirmed that no significance was found in the STOI results when processing the Volvo noise. However, the SE systems provided significantly different results for the Volvo noise when the PESQ, SSNR and LSD were the evaluation metrics of concern ($p = 5e\text{-}5$). Although the PESQ scores appear to be close, the differences between the averages of some groups, representing the different SE algorithms, are large enough to be statistically significant, resulting in a p-value of less than 4e-4.

The performance of all SEs deteriorated with diminishing SNR. Overall, the proposed DNN-QCM consistently provided better SE performance. This is then followed by the DNN-CRM and subsequently, the DNN-IRM methods. When compared to the DNN-IRM, the DNN-QCM led to an improvement of approximately 6.5% in STOI score, 11.0% in PESQ score, 35.7% in SSNR score and 28.1% in LSD score. When compared to the DNN-CRM, improvements of approximately 1.9% in STOI score, 3.3% in PESQ score, 1.2% in SSNR score and 4.8% in LSD score were obtained. From a statistical standpoint, the Tukey's HSD test [49] performed following the ANOVA confirmed that the DNN-QCM performed significantly better than the unprocessed speech, MMSE and NMF-based methods. However, although improved speech intelligibility and quality, and lower noise distortion were observed with the DNN-QCM in most conditions when compared to DNN-IRM and DNN-CRM, the differences in the evaluation metric scores were not statistically significant. The exception was where

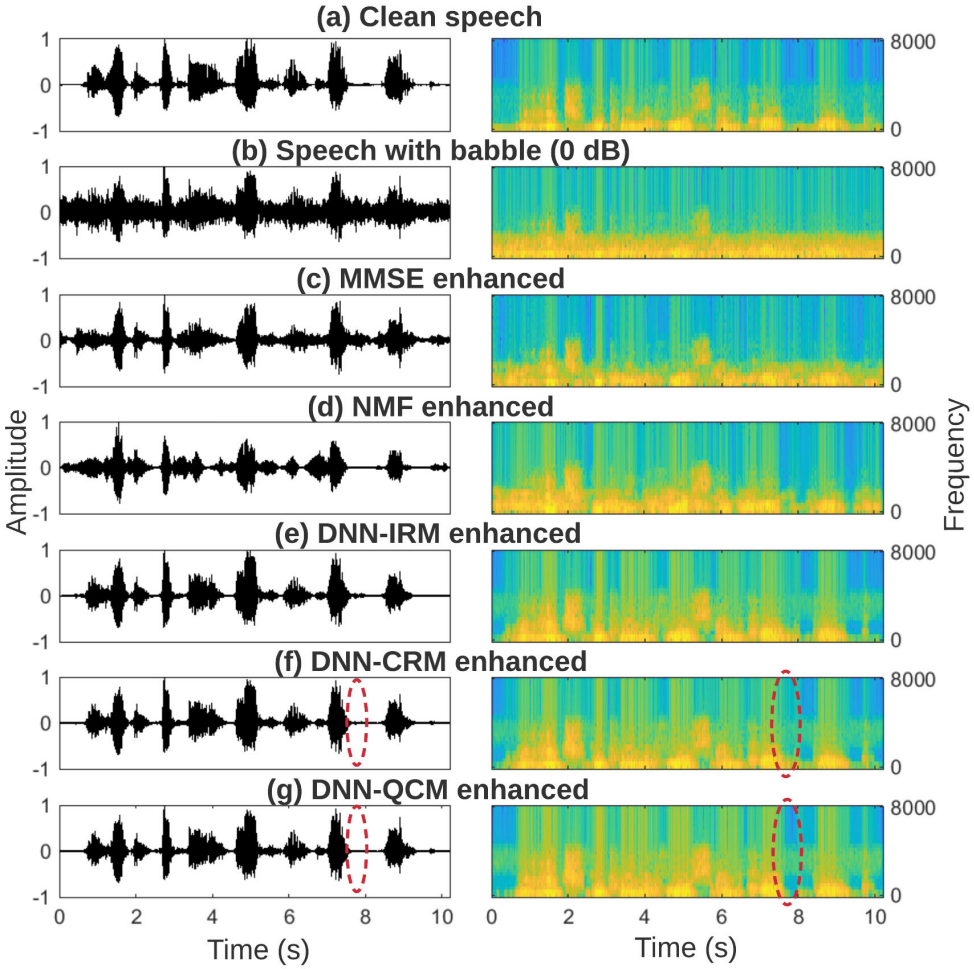


FIGURE 9. Example visualizations of the denoising performance of the different SE methods compared on a speech sample contaminated with babble noise at 0 dB. The circled portions depict an example segment of the speech where the DNN-QCM outperforms the DNN-CRM method in speech denoising.

the DNN-QCM provided significantly better STOI scores than the DNN-IRM. The DNN-CRM did not perform significantly better than the DNN-IRM in any conditions. This suggests that the DNN-based algorithms have reached near saturation in speech denoising in the conditions tested. Differences between the evaluation metric scores were deemed significant if the p-value obtained was smaller than 0.05.

Example visualizations of the enhanced signals obtained from the various SE systems are provided in Fig. 9. The figure shows example resulting time-domain waveforms of the enhanced signals and their corresponding spectrograms when speech contaminated with babble at 0 dB SNR was presented to the SE systems. The waveforms and spectrograms of the clean speech and noisy speech mixture (the unprocessed signal) have been additionally provided for comparison. The MMSE method removed much high-frequency noise components but it also removed some high frequency speech components as a result, leading to the introduction of distortion. It also performed worst in denoising silenced segments. The NMF method fell short of the DNN methods in terms of low-frequency noise suppression. The DNN methods were able to well segregate the beginning and end of the voiced fragments of the noisy speech mixture. The DNN-QCM method provided better high-frequency noise suppression than the DNN-IRM method and from observing the time-domain waveforms, it also excelled in suppressing low-amplitude noise. The DNN-QCM performed similarly to the DNN-CRM, however, some obvious improvements can be observed with the DNN-QCM. An example segment of the enhanced signals where the DNN-QCM can be seen to outperform the DNN-CRM has been marked with superimposed dashed circles in Fig. 9. Here, low-amplitude noise spanning across all frequencies were suppressed in the DNN-QCM but not in the DNN-CRM. In order to assess the generalization performance of the proposed SE method, noises that have been excluded from the training dataset (unseen noise types 'f16' and 'factory 2') were also fed to the SE for testing. The resulting outcomes are shown in Fig 10. The enhancement performance worsened for all

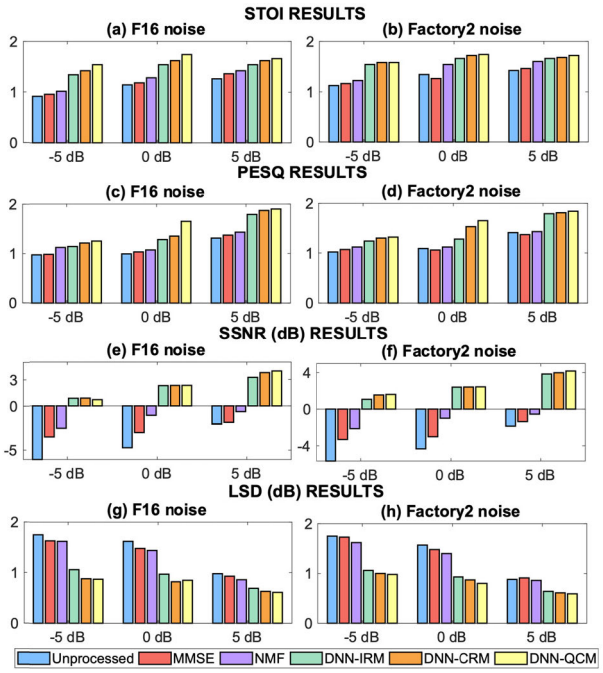


FIGURE 10. Evaluation performance comparisons of the various SE algorithms, including the proposed DNN-based SE with QCM, on untrained noise types, 'f16' and 'factory 2' noise.

SE algorithms when unseen noise types were presented. Nevertheless, a similar performance pattern to the seen noise test was obtained and the DNN with the proposed QCM continued to show better denoising capability than the other methods across both unseen noise types and at different SNR conditions.

D. COMPLEXITY, CONVERGENCE AND TIMING ANALYSIS OF THE SE NETWORKS

The training complexity of the DNN networks depends on the network parameters, forward-backward propagation for network tuning, quantity of neurons in the hidden layers and weights. A higher number of neurons leads to greater network complexity. In this paper, the number of layers and neurons in each layer (hidden as well as visible) for all the DNN-based methods (i.e., DNN-IRM, DNN-CRM and DNN-QCM) were kept the same. According to [20], by defining the dimension of the input acoustic features as X_i, number of training data points (size of the training target) as N_D, number of hidden layers as N_H, number of output neurons as N_O and number of epochs for parameter turning as N_E, the complexity of the DNNs employed can be quantified by the big O notation as $O(N_D N_E(X_i + N_H + 2N_H^2 + N_H N_O))$.

As shown in Figure 11, even though the same number of neurons and layers were used, the proposed DNN-QCM converged faster as a result of the quantization and DNAT. The RBM pre-training led to lower values of loss functions. The MSEs were observed to begin plateauing when the number of epochs reached approximately 50 and close to a full plateau was reached when the number of epochs is 100. This led to the decision to set the total number of epochs for training to 100, as mentioned earlier in Section III.

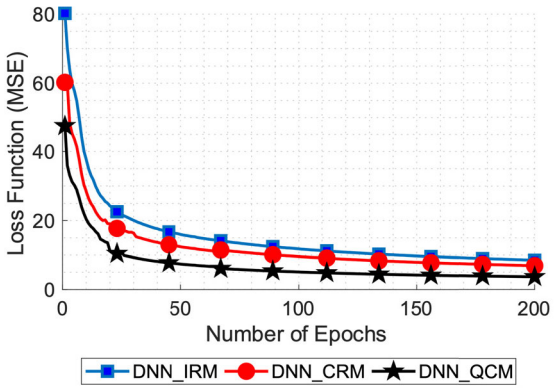


FIGURE 11. Convergence of training and validation errors obtained when tuning the parameters of the DNN-IRM, DNN-CRM and DNN-QCM.

TABLE 3. Training and inference time of the proposed DNN-QCM versus the DNN-IRM and DNN-CRM methods.

SE	Training time (ms)	Inference time (ms)
DNN-IRM	**1128**	41
DNN-CRM	1596	42
DNN-QCM	1379	**38**

The average training and inference time required for processing each frame of 25 ms were observed on a CPU for the DNN-QCM method to further evaluate the impact of using the sum tables NCC calculation method, quantization techniques and DNAT. These were compared to the DNN-IRM and the DNN-CRM, where NCCs were calculated in a conventional manner, and DNAT and quantization techniques were not implemented. As shown in Table 3, it took a longer time to train the DNN-QCM and DNN-CRM as their procedures involved calculating the NCCs for obtaining the required ICC factors used to adjust the IRM. With the introduction of the DNAT; alternative method of calculating the NCCs using sum tables; and quantization techniques to reduce the bit representations within the DNN; the training time of the proposed DNN-QCM was successfully reduced by approximately 15.7% when compared with the DNN-CRM. As for the inference time, which refers to the time required for a frame to undergo the SE processing, the DNN-IRM and DNN-CRM spent almost a similar amount of time for processing noisy speech. Quantization techniques have effectively reduced the inference time for the DNN-QCM due to reduced bit operations. When compared with the DNN-CRM, the inference time for the DNN-QCM was approximately 10.5% less.

V. CONCLUSION

Much room for improvement exists for monaural SE algorithms. They can be made more compact, computationally efficient and powerful in denoising under low SNR mixture and in non-stationary noise condition. In this paper, an efficient correlation-based ratio mask representation is proposed to address these limitations in SE algorithms. The proposed mask, coined as the QCM (quantized correlation mask), operates based on the CASA-DNN model for SE. In the QCM, adaptive ICC factors are used to adjust the ratio mask to more accurately retain and suppress speech and noise components. Since the traditional method of calculating NCCs, required to obtain the ICC factors, is computationally expensive, an alternative method of calculating NCCs using sum tables is presented. To increase noise robustness, a correlation-based DNAT (dynamic noise aware training) is proposed to be used in conjunction with the QCM. Quantization techniques are further applied to the QCM, neural network weights and acoustic features extracted to make the DNN more compact. With quantization, the weights within the employed neural network could be reduced to a 5-bit integer representation from a 32-bit float representation. If 100 different weights are employed in the DNN, this would mean achieving a compression rate of around 2. Despite the introduction of the DNAT, the sum tables and quantization techniques have led to approximately 15.7% and 10.5% reduction in the training and inference time respectively, for executing the pipeline of the proposed method. Furthermore, the proposed DNN-QCM method outperformed the reference methods, which included DNN-based SE systems trained with IRM and another correlation-based training target, and a model-based and NMF-based SE. Future work will include investigation of the DNN-QCM's denoising and generalization performance when trained with larger datasets having a wider range of SNRs, and implementation of dynamic quantization and real-time training capability to the DNN-QCM while keeping the memory and computational requirements at a minimum. The viability of the application of the proposed compact speech enhancement algorithm in hearing prostheses will also be assessed.

REFERENCES

[1] T. Van den Bogaert, S. Doclo, J. Wouters, and M. Moonen, "Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids," *J. Acoust. Soc. Amer.*, vol. 125, no. 1, pp. 360–371, Jan. 2009.

[2] L.-P. Yang and Q.-J. Fu, "Spectral subtraction-based speech enhancement for cochlear implant patients in background noise," *J. Acoust. Soc. Amer.*, vol. 117, no. 3, pp. 1001–1004, Mar. 2005.

[3] P. Sorqvist, P. Handel, and B. Ottersten, "Kalman filtering for low distortion speech enhancement in mobile communication," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1997, pp. 1219–1222.

[4] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5024–5028.

[5] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[6] H. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series, With Engineering Applications*. Cambridge, U.K.: Mit Press, 1949.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[8] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 2, pp. 137–145, Apr. 1980.

[9] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Comput. Speech Lang.*, vol. 8, no. 4, pp. 297–336, Oct. 1994.

[10] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, Sep. 2004.

[11] M. C. Anzalone, L. Calandruccio, K. A. Doherty, and L. H. Carney, "Determination of the potential benefit of time-frequency gain manipulation," *Ear Hearing*, vol. 27, no. 5, pp. 480–492, Oct. 2006.

[12] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *J. Acoust. Soc. Amer.*, vol. 125, no. 4, pp. 2336–2347, Apr. 2009.

[13] C. Hummersone, T. Stokes, and T. Brookes, "On the ideal ratio mask as the goal of computational auditory scene analysis," in *Signals and Communication Technology*. Berlin, Germany: Springer, 2014, pp. 349–368.

[14] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7092–7096.

[15] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[16] S. Liang, W. Liu, W. Jiang, and W. Xue, "The optimal ratio time-frequency mask for speech separation in terms of the signal-to-noise ratio," *J. Acoust. Soc. Amer.*, vol. 134, no. 5, pp. 452–458, Nov. 2013.

[17] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 3, pp. 483–492, Mar. 2016.

[18] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 708–712.

[19] T. Lan, Y. Lyu, W. Ye, G. Hui, Z. Xu, and Q. Liu, "Combining multi-perspective attention mechanism with convolutional networks for monaural speech enhancement," *IEEE Access*, vol. 8, pp. 78979–78991, 2020.

[20] N. Saleem, M. I. Khattak, M. Al-Hasan, and A. B. Qazi, "On learning spectral masking for single channel speech enhancement using feedforward and recurrent neural networks," *IEEE Access*, vol. 8, pp. 160581–160595, 2020.

[21] Y. Wang, G. Yu, J. Wang, H. Wang, and Q. Zhang, "Improved relativistic cycle-consistent GAN with dilated residual network and multi-attention for speech enhancement," *IEEE Access*, vol. 8, pp. 183272–183285, 2020.

[22] D. Takeuchi, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada, "Real-time speech enhancement using equilibriated RNN," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 851–855.

[23] S. Han, J. Pool, J. Tran, and W. J. Dally, "Learning both weights and connections for efficient neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Dec. 2015, pp. 1135–1143.

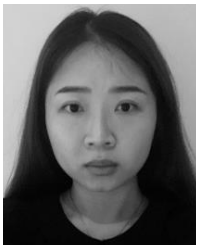
[24] C.-T. Liu, T.-W. Lin, Y.-H. Wu, Y.-S. Lin, H. Lee, Y. Tsao, and S.-Y. Chien, "Computation-performance optimization of convolutional neural networks with redundant filter removal," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 66, no. 5, pp. 1908–1921, May 2019.

[25] E. H. Lee, D. Miyashita, E. Chai, B. Murmann, and S. S. Wong, "LogNet: Energy-efficient neural networks using logarithmic computation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5900–5904.

[26] J.-Y. Wu, C. Yu, S.-W. Fu, C.-T. Liu, S.-Y. Chien, and Y. Tsao, "Increasing compactness of deep learning based speech enhancement models with parameter pruning and quantization techniques," *IEEE Signal Process. Lett.*, vol. 26, no. 12, pp. 1887–1891, Dec. 2019.

[27] J. H. Ko, J. Fromm, M. Philipose, I. Tashev, and S. Zarar, "Precision scaling of neural networks for efficient audio processing," in *Proc. 31st Conf. Neural Inf. Process. Syst.*, Dec. 2017, pp. 1–5.

[28] Y.-T. Hsu, Y.-C. Lin, S.-W. Fu, Y. Tsao, and T.-W. Kuo, "A study on speech enhancement using exponent-only floating point quantized neural network (EOFP-QNN)," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2018, pp. 566–573.
[29] X.-H. Chang and Y. Liu, "Robust $\mathcal{H}_\infty$ filtering for vehicle sideslip angle with quantization and data dropouts," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 10435–10445, Oct. 2020.
[30] J. Xiong, X. Chang, J. H. Park, and Z. Li, "Nonfragile fault-tolerant control of suspension systems subject to input quantization and actuator fault," *Int. J. Robust Nonlinear Control*, vol. 30, no. 16, pp. 6720–6743, Jul. 2020.
[31] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallet, and N. L. Dahlgren, *DARPA-TIMIT: Acoustic-Phonetic Continuous Speech Corpus*. Washington, DC, USA: US Department of Commerce, 1993.
[32] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993.
[33] F. Bao and W. H. Abdulla, "A new ratio mask representation for CASA-based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 1, pp. 7–19, Jan. 2019.
[34] J. Luo and E. E. Konofagou, "A fast normalized cross-correlation calculation method for motion estimation," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 57, no. 6, pp. 1347–1357, Jun. 2010.
[35] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
[36] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2001, pp. 749–752.
[37] J. H. L. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. 5th Int. Conf. Speech Lang. Process.*, Sydney, NSW, Australia, vol. 7, Dec. 1998, pp. 2819–2822.
[38] P. J. Wolfe and S. J. Godsill, "Towards a perceptually optimal spectral amplitude estimator for audio signal enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Aug. 2000, pp. 821–824.
[39] Y. Wang, K. Han, and D. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 21, no. 2, pp. 270–279, Feb. 2013.
[40] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.
[41] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.
[42] G. E. Hinton, "A practical guide to training restricted Boltzmann machines," in *Neural Networks: Tricks Trade*. Berlin, Germany: Springer, 2012, pp. 599–619.
[43] R. Malouf, "A comparison of algorithms for maximum entropy parameter estimation," in *Proc. 6th Conf. Natural Lang. Learn. COLING*, Aug. 2002, pp. 49–55.
[44] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7398–7402.
[45] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.
[46] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn.*, vol. 12, pp. 1–39, Jul. 2011.
[47] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 4266–4269.
[48] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 10, pp. 2140–2151, Oct. 2013.
[49] J. W. Tukey, "Comparing individual means in the analysis of variance," *Biometrics*, vol. 5, no. 2, pp. 99–114, 1949.



SALINNA ABDULLAH (Graduate Student Member, IEEE) received the M.Eng. degree in electronic engineering from the Department of Computer Science, University College London (UCL), London, U.K., in 2017, where she is currently pursuing the Ph.D. degree with the Analog and Biomedical Electronics Group, Department of Electrical and Electronic Engineering. Her research interests include FPGA design, efficient application of speech enhancement, image processing, and deep-learning methods in wearable devices. She was a recipient of the EPSRC Industrial Strategy Studentship. She was awarded the Cisco Prize for Most Outstanding Female Engineer during her final year of undergraduate study.

MAJID ZAMANI (Member, IEEE) was born in Tehran, Iran, in 1984. He received the M.Sc. degree in microelectronics from the Islamic Azad University, Science, and Research Branch, Tehran, in 2011, and the Ph.D. degree from the University College London (UCL), London, U.K., in 2017. He is currently a Research Associate with the Analog and Biomedical Electronics Group, UCL. His research interests include design and fabrication of advanced and energy-efficient computational systems utilizing pattern recognition, machine learning, and computer vision algorithms, especially for wearable and implantable biomedical applications. He was a recipient of the Overseas Research Scholarship and the UCL Graduate Research Scholarship to pursue his Ph.D. degree. He was also a recipient of the Best Researcher M.Sc. Student Award.

ANDREAS DEMOSTHENOUS (Fellow, IEEE) received the B.Eng. degree in electrical and electronic engineering from the University of Leicester, Leicester, U.K., in 1992, the M.Sc. degree in telecommunications technology from Aston University, Birmingham, U.K., in 1994, and the Ph.D. degree in electronic and electrical engineering from University College London (UCL), London, U.K., in 1998.

He is currently a Professor with the Department of Electronic and Electrical Engineering, UCL, and also leads the Analog and Biomedical Electronics Group. He has made outstanding contributions to improving safety and performance in integrated circuit design for active medical devices, such as spinal cord and brain stimulators. He has numerous collaborations for cross-disciplinary research, both within the U.K., and internationally. He has authored more than 300 articles in journals and international conference proceedings, several book chapters, and holds several patents. His research interests include analog and mixed-signal integrated circuits for biomedical, sensor, and signal processing applications. He is a Fellow of the Institution of Engineering and Technology and a Chartered Engineer. He was a co-recipient of a number of Best Paper awards and has graduated many Ph.D. students. He was an Associate Editor from 2006 to 2007 and the Deputy Editor-in-Chief from 2014 to 2015 of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—II: EXPRESS BRIEFS, and an Associate Editor from 2008 to 2009 and the Editor-in-Chief from 2016 to 2019 of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: REGULAR PAPERS. He is an Associate Editor of the IEEE TRANSACTIONS ON BIOMEDICAL CIRCUITS AND SYSTEMS and serves on the International Advisory Board of Physiological Measurement. He has served on the technical committees for a number of international conferences, including the European Solid-State Circuits Conference (ESSCIRC) and the International Symposium on Circuits and Systems (ISCAS).