A hybrid noise suppression filter for accuracy enhancement of commercial speech recognizers in varying noisy conditions

Kit Yan Chan[1], Pei Chee Yong[1], Sven Nordholm[1], Cedric K.F. Yiu[2], Hak Keung Lam[3]

[1]Department of Electrical and Computer Engineering, Curtin University, Perth, Australia

[2]Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong

[3]Department of Informatics, King's College London, United Kingdom

Abstract —

Commercial speech recognizers have made possible many speech control applications such as wheelchair, tone-phone, multifunctional robotic arms and remote controls, for the disabled and paraplegic. . However, they have a limitation in common in that recognition errors are likely to be produced when background noise surrounds the spoken command, thereby creating potential dangers for the disabled if recognition errors exist in the control systems. In this paper, a hybrid noise suppression filter is proposed to interface with the commercial speech recognizers in order to enhance the recognition accuracy under variant noisy conditions. It intends to decrease the recognition errors when the commercial speech recognizers are working under a noisy environment. It is based on a sigmoid function which can effectively enhance noisy speech using simple computational operations, while a robust estimator based on an adaptive-network-based fuzzy inference system is used to determine the appropriate operational parameters for the sigmoid function in order to produce effective speech enhancement under variant noisy conditions. The proposed hybrid noise suppression filter has the following advantages for commercial speech recognizers: i) it is not possible to tune the inbuilt parameters on the commercial speech recognizers in order to obtain better accuracy; ii) existing noise suppression filters are too complicated to be implemented for real-time speech recognition; and iii) existing sigmoid function based filters can operate only in a single-noisy condition, but not under varying noisy conditions. The performance of the hybrid noise suppression filter was evaluated by interfacing it with

a commercial speech recognizer, commonly used in electronic products. Experimental results show that improvement in terms of recognition accuracy and computational time can be achieved by the hybrid noise suppression filter when the commercial recognizer is working under various noisy environments in factories.
***Index Terms* —** Fuzzy neural networks, noise suppression filter, ANFIS, speech recognition, commercial speech recognizer, sigmoid filter, speech enhancement

I. INTRODUCTION

Well-established speech recognition technologies benefit many rehabilitation and biomedical engineering industries in the development of health or assistance devices for paraplegics and the disabled, where speech recognition is used as a patient-machine interface which transfers control commands from patient to the machine and provides feedback from the machine to the patient. Although advanced sensors for human movements such as eyes and tongue switches have been included in health or assistance devices, such sensing approaches have many limitations which make these devices less efficient for the user. In fact, speech control is simpler in interfacing between the patients and the assistance devices, and it has been implemented for the control of server-assistance devices for the disabled [44, 45]. In the commercial and industrial sectors, speech controls [2, 27] are used in factory automation [32], warehouse automation [1], and industrial robotic control [26]. Disabled people can give verbal commands or input data to control the manufacturing system without the necessity for physical contact since speech is the only thing required to control the manufacturing systems [24].

During the development of the commercial recognizer mechanism, much research regarding speech recognition was conducted on enhancing speech recognition accuracy by developing effective recognition algorithms [23, 14], identification mechanisms for capturing significant speech features for recognition [30, 20], recognition algorithms for distorted and contaminated speech [42], etc. However, these approaches have a common limitation in that the development of speech recognizers has been based only on a database which consists of a limited number of speech signals contaminated by certain types of acoustic noise. It is

impractical for industrial sectors to manufacture a commercial speech recognizer which can work ideally in every noisy environment, due to the limitations of cost and time. Hence, conventional commercial recognizers can work accurately only under the trained noisy conditions, but might not work under untrained noisy conditions in which inaccurate recognitions are likely to occur. Recognition errors in a control system would create potential risks and dangers for the disabled user. Therefore, it is necessary to decrease the recognition errors for commercial speech recognizers that are working in noisy environments.

Although multi-channel beamformers [9, 15, 41] can be used to enhance noisy speech which is contaminated by near-end noise, such approaches have the commonly limitation that the signal source needs to be tracked, based on the difference between the signal spectrums collected from multiple channels. Also, their mechanisms are computationally complex and their inbuilt parameters need to be calibrated with respect to the location of noise sources and speech sources.

A practical way to improve recognition accuracy is to interface the commercial speech recognizer with a noise suppression filter which works effectively under multi-noise conditions for varying noisy environments. Noise suppression filters first identify both active and inactive periods of noisy speech, and then estimate the noise spectrum based on the inactive periods. Hence, an enhanced speech spectrum can be produced by removing the noise spectrum from the original speech spectrum [4]. However, several acoustic criteria need to be addressed in order to enhance the accuracy of commercial speech recognition, as annoying residual noise or musical noise can be perceivable when the noise spectrum is under-estimated, and original speech can be distorted leading to a loss of speech quality and intelligibility when the noise spectrum is over-estimated [29]. It is difficult to optimize those acoustic criteria even under a single condition with respect to a single user and a single noisy environment, as nonlinearities exist in those acoustic criteria; moreover, multi-objective optimization for satisfying those acoustic criteria needs to be handled. It is much more difficult to satisfy multi-noise conditions which involve multi-users and varying noisy environments.

Although much progress has been made in the development of noise suppression filters using different cost functions engaged with different acoustic criteria which are perceptual, intelligibility and quality [6, 3], these

approaches are often computationally complex and difficult to implement in real-case scenarios. Even though a powerful processor can be used to implement such complex filters, speech quality can only be improved with respect to the specified but limited acoustic criteria which may not be related to accuracy of speech recognition. Hence, they still result in poor speech recognition performance.

Hu et al. [16] show that sigmoid filters with low complexity can be implemented for real-time speech recognitions. These simple sigmoid filters overcome the limitation of the noise suppression filters which are too computationally complex for real-time implementation. To further increase the flexibility of speech enhancement for varying noise conditions, Yong et al. [39] have recently developed an advanced sigmoid gain function, which combines a logistic function with a hyperbolic tangent function. The advanced sigmoid gain function provides several filter parameters that can be adjusted to flexibly model exponential distributions, in order to achieve a balanced trade-off between many acoustic criteria such as noise reduction, speech distortion and musical noise [41]. With this trade-off, the accuracy of the commercial recognizer can be enhanced under a single-noisy condition. However, a particular set of filter parameters can be optimized only with respect to a single condition with a particular noise power level. Hence, it is necessary to adjust the filter parameters to maintain the trade-off under varying noisy conditions.

In this paper, a hybrid noise suppression filter, namely ANFIS-SF, is proposed based on the mechanisms of the ANFIS and the sigmoid filter, in order to improve the accuracy of the commercial recognizer operating in varying noisy conditions. To develop ANFIS-SF, a speech recognition problem is formulated in order to optimize the accuracy of the commercial recognizer with respect to a single-noisy condition. A global optimization method, namely particle swarm optimization (PSO), is used to initialize a set of optimal filter parameters, each of which is optimized with respect to the speech recognition problem, as PSO can effective in solving optimization problems with similar landscapes which are discontinuous, vastly multimodal and non-differentiable [25, 36, 37, 43]. Based on these optimal filter parameters for single-noise conditions, a robust estimator [18], ANFIS, is used to perform a mapping relationship between filter parameters and various noisy conditions, as this mapping relationship is highly nonlinear and ANFIS is an effective method

for nonlinear mapping [31, 22, 10, 13]. As the ANFIS provides appropriate filter parameters for the sigmoid filter with respect to varying noisy conditions, the resulting ANFIS-SF is likely to work effectively under such conditions.

The effectiveness of the ANFIS-SF is demonstrated by interfacing it with a commercial speech recognizer which is used in electronic products [28]. When compared with other existing noise suppression filters, ANFIS-SF produces better results in terms of recognition accuracy and computational time in factory environments.

II. ENHANCEMENT OF RECOGNITION ACCURACY UNDER MULTI-CONDITIONS

A commercial recognizer, $\Re(.)$, is designed to identify *n* inbuilt speech commands, $\{u^1, u^2, ..., u^n\}$ of which those speech commands can be single words such as numerical digits, 'yes' or 'no' decisions, 'left' or 'right' directions etc, or those speech commands can also be phrases, such as operational commands for manufacturing processes, speech controls for toys or audio players etc.

Let the *t*-th sample of the noisy speech, $x_j^i(t)$, received by $\Re(.)$ be denoted as,

$$x_j^i(t) = s_j^i(t) + v(t), \qquad i = 1,2,...,n \tag{1}$$

where $s_j^i(t)$, is the *i*-th speech command voiced out by the *j*-th regular user, with *j*=1,2..,*N* of which *N* is the number of regular users; $v(t)$, is the background noise; $t = 1,2,...,m$; and *m* is the number of samples. Let $\hat{i}$ be the recognized command from $\Re(.)$ for the noisy speech, $x_j^i(t)$, which is given by,

$$\hat{i} = \Re\left(x_j^i\right), \tag{2}$$

where $x_j^i = \left[x_j^i(1), x_j^i(2), ..., x_j^i(m)\right]$. If $\hat{i} = i$, a correct recognition is obtained with respect to x_j^i. Otherwise, an incorrect recognition occurs, if $\hat{i} \neq i$. When the power of $v(t)$ is large, recognition errors are likely to be produced by $\Re(.)$. A sigmoid filter, namely $G_{\text{SIG}}\left(\omega, \ell, \bar{\kappa}(\sigma)\right)$ with three filter parameters,

$\bar{\kappa}(\sigma)=\left[\kappa_1(\sigma),\kappa_2(\sigma),\kappa_3(\sigma)\right]$, can be used to enhance the accuracy of $\Re(.)$, where ω is a real angular center frequency given by $\omega\in\left[\omega_{0,}\omega_{1,}....,\omega_{K-1}\right]$; ℓ is the time frame index given by $\ell\in[0,1,....,L-1]$; K is the number of bands; L is the number of frames; and σ is the estimated degrees of signal to noise ratio (SNR).

Using $G_{\mathrm{SIG}}\left(\omega,\ell,\bar{\kappa}(\sigma)\right)$, the estimate of clean speech spectrum, $\hat{S}_j^i(\omega,\ell)$, with respect to $s_j^i(t)$ can be obtained by:

$$\hat{S}_j^i(\omega,\ell)=G_{\mathrm{SIG}}\left(\omega,\ell,\bar{\kappa}(\sigma)\right)\cdot X_j^i(\omega,\ell) \tag{3}$$

where $X_j^i(\omega,\ell)$ is the M-point short time Fourier transformation of $x_j^i(t)$; and $G_{\mathrm{SIG}}\left(\omega,\ell,\bar{\kappa}(\sigma)\right)$ is formulated as:

$$\begin{aligned}&G_{\mathrm{SIG}}\left(\omega,\ell,\bar{\kappa}(\sigma)\right)=G_{\mathrm{SIG}}\left(\omega,\ell,\kappa_1(\sigma),\kappa_2(\sigma),\kappa_3(\sigma)\right)\\&=\frac{1}{1+e^{-\kappa_1(\sigma)\cdot\left(\hat{\xi}(\omega,\ell)-\kappa_2(\sigma)\right)}}\cdot\frac{1-e^{-\kappa_3(\sigma)\cdot\hat{\xi}(\omega,\ell)}}{1+e^{-\kappa_3(\sigma)\cdot\hat{\xi}(\omega,\ell)}}\quad;\end{aligned} \tag{4}$$

$\hat{\xi}(\omega,\ell)$ is the estimate of a priori SNR, which is determined based on the modified decision directed approach [39] and is given by

$$\hat{\xi}(\omega,\ell)=\beta\frac{\left|G_{\mathrm{SIG}}\left(\omega,\ell-1,\bar{\kappa}(\sigma)\right)\cdot X(\omega,\ell)\right|^2}{\hat{\lambda}_v(\omega,\ell)}+(1-\beta)P\left(\gamma(\omega,\ell)-1\right)$$

where $\hat{\lambda}_v(\omega,\ell)$ and $G_{\mathrm{SIG}}\left(\omega,\ell-1,\bar{\kappa}(\sigma)\right)$ denote, respectively, the estimated noise PSD from [38] and the gain value from the preceding frame. The parameter β denotes the smoothing factor, $P(.)$ denotes the half-wave rectification and $\gamma(\omega,\ell)$ denotes the a posteriori SNR. With $\hat{\xi}(\omega,\ell)$, σ can be determined as $\sigma=E\left(\hat{\xi}(\omega,\ell)\right)$.

In $G_{\mathrm{SIG}}\left(\omega,\ell,\bar{\kappa}(\sigma)\right)$, $\bar{\kappa}(\sigma)$ needs to be optimized with respect to σ. In the first term of $G_{\mathrm{SIG}}\left(\omega,\ell,\bar{\kappa}(\sigma)\right)$ illustrated in (4), the sigmoid slope of the logistic function can be adjusted in order to maximize sensitive

towards speech and minimize sensitive towards variation of noise, by tuning $\kappa_1(\sigma)$ and $\kappa_2(\sigma)$ with respect to σ. To further enhance the effectiveness of noise reduction at low SNR, $\kappa_3(\sigma)$ can be used to make $G_{\mathrm{SIG}}\left(\omega,\ell,\bar{\kappa}(\sigma)\right)$ trending to be zero under a very low σ. Therefore, the behavior of $G_{\mathrm{SIG}}\left(\omega,\ell,\bar{\kappa}(\sigma)\right)$ can be made similar to different conventional noise filters by optimizing $\bar{\kappa}(\sigma)=\left[\kappa_1(\sigma),\kappa_2(\sigma),\kappa_3(\sigma)\right]$ with respect to σ [39].

When working under a single condition, the filter parameters, $\bar{\kappa}(\sigma)$, need to be optimized only with respect to a single σ, by solving the following optimization problem in order to maximize the accuracy of $\Re(.)$:

$$J_s\left(\bar{\kappa}(\sigma)\right)=\max_{\bar{\kappa}(\sigma)}\sum_{j=1}^{N}\sum_{i=1}^{n}\theta_j^i \ , \tag{5}$$

subject to

$$\theta_j^i=\begin{cases}0 & \text{if } \Re\left(\hat{s}_j^i\right)\neq i, \text{ which is incorrect recognition}\\ 1 & \text{if } \Re\left(\hat{s}_j^i\right)=i, \text{ which is correct recognition}\end{cases},$$

where $\hat{s}_j^i$ is the time domain of $\hat{S}_j^i(\omega,\ell)$ in (3) which is obtained by the inverse *M*-point short-time Fourier transformation of $\hat{S}_j^i(\omega,\ell)$. However, it is impractical to develop a filter which can only work effectively with respect to a single condition, as the power of noise and the power of user speeches cannot be a constant in an operational environment for the commercial recognizer. Hence, multi-conditions need to be addressed to generate a filter which can enhance accuracy of the commercial recognizer under multiple SNR. For multi-conditions, the following optimization problem is solved, in order to maximize accuracy of $\Re(.)$ under varying σ:

$$J_M\left(\bar{\mathrm{K}}\right)=\sum_{k=1}^{p}J_s\left(\bar{\kappa}(\sigma_k)\right), \tag{6}$$

where all σ_k are within the specified minimum and maximum of SNRs with respect to the operational environments for the commercial recognizer; and $\bar{\mathrm{K}}^{opt} = \left[\bar{\kappa}^{opt}(\sigma_1), \bar{\kappa}^{opt}(\sigma_2), \ldots \ldots, \bar{\kappa}^{opt}(\sigma_p) \right]$ is the filter parameter set for multi-conditions of which $\bar{\kappa}^{opt}(\sigma_k)$ is optimal with respect to the SNR of σ_k; and p is the number of multi-conditions with multiple SNR. Since the optimization problem formulated in (6) involved the step function which represents the correct recognitions, it is non-differentiable. Hence, local search methods, which require gradient information to trace optimum, are not appropriate for solving this optimization problem. Since the particle swarm optimization performs effectively on solving similar optimization problems which are discontinuous, vastly multimodal and non-differentiable [25, 36, 37, 43], the particle swarm optimization (discussed in Section IIIA) is used to solve this optimization problem.

III. HYBRID ANFIS AND SIGMOID FILTER

In this paper, a hybrid filter namely, ANFIS-SF, is proposed based on the mechanisms of the ANFIS and the sigmoid filter, in order to enhance the accuracy of the commercial recognizer for multi-conditions formulated in (6). To develop ANFIS-SF, PSO is first used to initialize a set of optimal filter parameters, each of which is optimized with respect to a single condition formulated in (5). Then an ANFIS is developed based on these optimal filter parameters, in order to create a map between the filter parameters and the SNR.

Figure 1 illustrates the mechanism of the ANFIS-SF. The filter parameters are first determined by the ANFIS based on the degree of estimate of SNR, which is within the operational environments for multi-conditions. Then, the sigmoid filter enhances the noisy speech based on the determined filter parameters. Correct recognitions are more likely to be produced by the commercial recognizer based on the enhanced speeches. Hence, the multi-conditions formulated in (6) can be addressed. Detailed descriptions of the PSO and the ANFIS are discussed in the following two sub-sections.

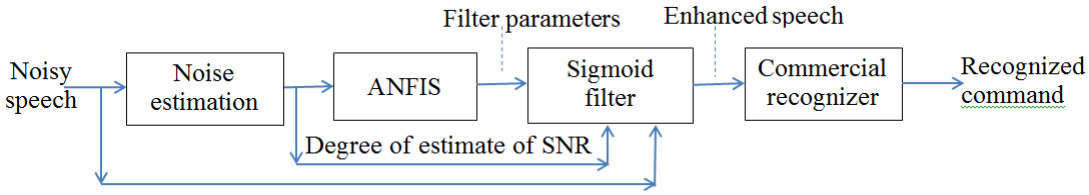

Figure 1 Mechanism of the ANFIS-SF

A. Initiation of filter parameters for single conditions using PSO

The PSO consists of N_s particles, where the position of the *j*-th particle at the *g*-th generation is represented by:

$$P_j^g(\sigma_i) = \left(\kappa_{j,1}^g(\sigma_i), \kappa_{j,2}^g(\sigma_i), \kappa_{j,3}^g(\sigma_i)\right), \tag{7}$$

which are the filter parameters used in the sigmoid filter, $G_{\mathrm{SIG}}\left(\omega, \ell, P_j^g(\sigma_i)\right)$, against a single condition for a particular SNR, σ_i. At the 1-*st* generation with *g*=1, all $\kappa_{j,k}^g(\sigma_i)$, with *k*=1, 2, 3, are generated randomly within their operational ranges, given as $\kappa_{j,1}^g(\sigma_i), \kappa_{j,2}^g(\sigma_i) \in [0..1]$ and $\kappa_{j,3}^g(\sigma_i) \in [0..15]$, which can be referred to [39]. When g>1, each $\kappa_{j,k}^g(\sigma_i)$ are updated based on its velocity, $vel_{j,k}^g(\sigma_i)$, by the following formulation (8):

$$\kappa_{j,k}^g(\sigma_i) = \kappa_{j,k}^{g-1}(\sigma_i) + vel_{j,k}^g(\sigma_i) \tag{8}$$

where

$$\begin{aligned} vel_{j,k}^g(\sigma_i) = C\Big(& v_{j,k}^{g-1}(\sigma_i) + \varphi_1 \cdot \gamma \cdot \left(pbest_{j,k}(\sigma_i) - \kappa_{j,k}^{g-1}(\sigma_i)\right) \\ & + \varphi_2 \cdot \gamma \cdot \left(gbest_k(\sigma_i) - \kappa_{j,k}^{g-1}(\sigma_i)\right)\Big) \end{aligned} \tag{9}$$

$pbest_j(\sigma_i) = \left[pbest_{j,1}(\sigma_i), pbest_{j,2}(\sigma_i), pbest_{j,3}(\sigma_i)\right]$, and

$gbest(\sigma_i) = \left[gbest_1(\sigma_i), gbest_2(\sigma_i), gbest_3(\sigma_i)\right]$;

$pbest_j(\sigma_i)$ denotes the best previous position of a particle recorded from the previous generation; $gbest(\sigma_i)$ denotes the position of the best particle among all particles; γ denotes a random number in the range of [0,1]; *w* is an inertia weight factor; ϕ_1 and ϕ_2 are the acceleration constants [11]; and *C* denotes the constriction factor, that ensures the PSO converges [5], which is given by:

$$C = \frac{2}{\left|2-\phi-\sqrt{\phi^2-4\phi}\right|}, \text{ with } \phi=\phi_1+\phi_2 \text{ and } \phi>4. \tag{10}$$

The PSO utilizes $pbest_j(\sigma_i)$ and $gbest(\sigma_i)$ to modify the current location of all $\kappa_{j,k}^{g}(\sigma_i)$ in order to prevent them from moving in the same direction, but to converge gradually towards $pbest_i$ and *gbest* [12]. To further refine the dynamic of $\kappa_{j,k}^{g}(\sigma_i)$, $vel_{j,k}^{g}(\sigma_i)$ is limited by a value which was set as 10%–20% of its range. This limit is employed to avoid $\kappa_{j,k}^{g}(\sigma_i)$ from flying past good solutions or exploring insufficient local solutions. The searching process of the PSO stops when it converges to a satisfactory filter parameter with respect to (5), where the satisfactory filter parameters are denoted as $\bar{\kappa}^{opt}(\sigma_i)=\left[\kappa_{j,1}^{opt}(\sigma_i),\kappa_{j,2}^{opt}(\sigma_i),\kappa_{j,3}^{opt}(\sigma_i)\right]$, which can address the enhancement of speech recognition with respect to a single condition formulated in (5).

To address the multi-conditions formulated in (6), the PSO can be run for *p* times where each run targets a particular SNR condition, σ_i with *i*=1, 2, …, *p*. However, it is impractical to find large sets of filter parameters in which each set of filter parameters is optimal with respect to only a single SNR. Hence, the satisfactory filter parameter set, $\bar{\mathrm{K}}=\left[\bar{\kappa}^{opt}(\sigma_1),\bar{\kappa}^{opt}(\sigma_2),...,\bar{\kappa}^{opt}(\sigma_p)\right]$, determined by the PSO is used to develop an ANFIS in order to construct an input-output mapping between σ and $\bar{\kappa}'(\sigma)$. A description of the development of an ANFIS is given in the following subsection.

B. ANFIS *for multi-conditions*

Although simple neural networks are capable of modelling nonlinear systems, the models generated are implicit. It is difficult to analyze the relationship between filter parameters and SNR within the simple neural network. Therefore, ANFIS is proposed, whereby a non-linear and explicit model can be generated in order to represent the relationship between filter parameters and SNR. In ANFIS, the architecture of Takagi-Sugeno fuzzy model [34] with one input and three outputs is used, where the input represents the SNR, σ, and the outputs represents the estimates of the three filter parameters which are denoted by,

$$\hat{\kappa}(\sigma)=\left[\hat{\kappa}_1(\sigma),\hat{\kappa}_2(\sigma),\hat{\kappa}_3(\sigma)\right]. \tag{11}$$

Using the mechanism of ANFIS, $\hat{\kappa}(\sigma)$ can be estimated with respect to any σ, where the ANFIS consists of n_{rule} fuzzy rules, and the *g*-th fuzzy rule is given by:

$$R_g:\ \text{IF } \sigma \text{ is } A_g(\sigma) \text{ THEN } z_g^j = w_{0,g} + \sigma \cdot w_{1,g}^j \text{ with } j=1,2,3; \tag{12}$$

$g=1,2,....,n_{rule}$; $w_{0,g}^j$ and $w_{1,g}^j$ are the consequent coefficients with respect to the polynomial of the *g*-th rule and the estimate of the *j*-th filter parameter, $\hat{\kappa}_j(\sigma)$; z_g^j is the consequence of the polynomial with respect to the *g*-th rule and $\hat{\kappa}_j(\sigma)$; $A_g(\sigma)$ denotes the membership function of the *g*-th fuzzy rule, which is given by:

$$A_g(\sigma)=\exp\left(\frac{-(\sigma-\mu_g)^2}{\rho_g}\right); \tag{13}$$

and μ_g and ρ_g are the mean and the variance of $A_g(\sigma)$ respectively. When the SNR is σ, the estimate of the *j*-th filter parameter, $\hat{\kappa}_j(\sigma)$,is given by:

$$\hat{\kappa}_j(\sigma)=\frac{\sum_{g=1}^{n_{rule}}\left(A_g(\sigma)\cdot\left(w_{0,g}^j+\sigma\cdot w_{1,g}^j\right)\right)}{\sum_{g=1}^{n_{rule}}\left(A_g(\sigma)\right)}. \tag{14}$$

Here, a widely used approach with fast convergence, namely Jang's algorithm [18], is used to determine the ANFIS parameters, $w_{0,g}^j$, $w_{1,g}^j$, μ_g, and ρ_g. The filter parameters, $\bar{\kappa}^{opt}(\sigma_i)=\left[\kappa_{j,1}^{opt}(\sigma_i),\kappa_{j,2}^{opt}(\sigma_i),\kappa_{j,3}^{opt}(\sigma_i)\right]$,

with i=1, 2,…, p, which are determined based on the PSO, are used to train the ANFIS, where that set of filter parameters is divided into two sets, namely training set, $\overline{\mathrm{T}}$, and validation set, $\overline{\Omega}$, such that $\bar{\kappa}^{opt}\left(\sigma_i\right) \in \overline{\mathrm{T}}$ and $\bar{\kappa}^{opt}\left(\sigma_j\right) \in \overline{\Omega}$, with all $i, j \in [1,2,...,p]$ but all $i \neq j$.

The training error, $e_{\overline{\mathrm{T}}}$, and the validation error, $e_{\overline{\Omega}}$, with respect to $\overline{\mathrm{T}}$ and $\overline{\Omega}$ are given by:

$$e_{\overline{\mathrm{T}}} = \sum_{\forall \bar{\kappa}^{opt}(\sigma_i) \in \overline{\mathrm{T}}} e_{ARE}\left(\bar{\kappa}^{opt}\left(\sigma_i\right)\right) \tag{15}$$

and $e_{\overline{\Omega}} = \sum_{\forall \bar{\kappa}^{opt}(\sigma_i) \in \overline{\Omega}} e_{ARE}\left(\bar{\kappa}^{opt}\left(\sigma_i\right)\right)$ respectively, (16)

where $e_{ARE}\left(\bar{\kappa}^{opt}\left(\sigma_i\right)\right)$ is the absolute difference between the real parameter, $\bar{\kappa}^{opt}\left(\sigma_i\right)$, and the estimated parameter, $\hat{\kappa}\left(\sigma_i\right)$, and $e_{ARE}\left(\bar{\kappa}^{opt}\left(\sigma_i\right)\right)$ is determined by:

$$e_{ARE}\left(\bar{\kappa}^{opt}\left(\sigma_i\right)\right) = e_{ARE}\left(\left[\kappa_1^{opt}\left(\sigma_i\right), \kappa_2^{opt}\left(\sigma_i\right), \kappa_3^{opt}\left(\sigma_i\right)\right]\right) = \sum_{j=1}^{3} \left| \frac{\kappa_j^{opt}\left(\sigma_i\right) - \hat{\kappa}_j\left(\sigma_i\right)}{\kappa_j^{opt}\left(\sigma_i\right)} \right| \tag{17}$$

In the early training stage for the ANFIS, both training errors and validation errors decrease gradually, as the characteristics of both training and test sets are usually similar. After both errors decrease to a certain level, the ANFIS becomes over-trained by the training set and the validation error increases with the decreased training error. During this time, the training of the ANFIS is terminated when the validation error starts to increase. Hereby, over-fitting can be avoided and the determined filter parameters, $\bar{\kappa}(\sigma)$, can be used as the filter parameters of the sigmoid filter. Hence, the multi-conditions formulated in (6) with varying σ can be addressed.

IV. EVALUATION OF ANFIS-SIGMOID FILTER

A. Implementation of a commercial recognizer

In this research, a commercial recognizer, namely the RSC3X synthesis microcontroller [28], is used to evaluate the effectiveness of the proposed ANFIS-SF since the commercial recognizer has been applied widely to various electronic products with speech control functions.

The commercial recognizer is a finite state machine, which can be in either a standby state or trigger state. When there is conversational interchange between the commercial recognizer and the user, the commercial recognizer is in the trigger state. Otherwise, the commercial recognizer is in the standby state where it is assumed that no conversational interchange is being conducted. During the trigger state, the trigger phrase voiced out by the user is passed to its inbuilt hidden Markov model which generates a set of likelihoods for the trigger phrase with respect to the speech commands. The speech command with the highest likelihood is the outcome of the commercial recognizer which is the recognized command. More than 99% accuracy can be produced when the commercial recognizer works under high SNRs [28], but wrong recognitions are likely to be produced under low SNRs.

The commercial recognizer was configured as in Figure 4, which simulates the environment of a noisy factory. Here the commercial recognizer was placed in front of a warehouse operator and a noise source was produced by a factory machine located beside the warehouse operator. The proposed ANFIS-SF was interfaced with the commercial recognizer in order to enhance the noisy speech before performing speech recognition. The commercial recognizer was implemented for the assigned machine, which can assign the five Christmas carol products including '*Jingle Bells*', '*Santa Claus is Coming to Town*', '*Sleigh Ride*', '*Let It Snow*', and '*Winter Wonderland*', where the five carol names are the speech commands embedded in the commercial recognizer. These are typical commands comprising phrases which can be used as the statements for commanding the assigned machine to deliver different song products. The assigned machine can deliver the correct song product if the speech command voiced by the factory operator is recognized correctly. If a

wrong command is recognized, the assigned machine either deliver correct song product or generates no action.

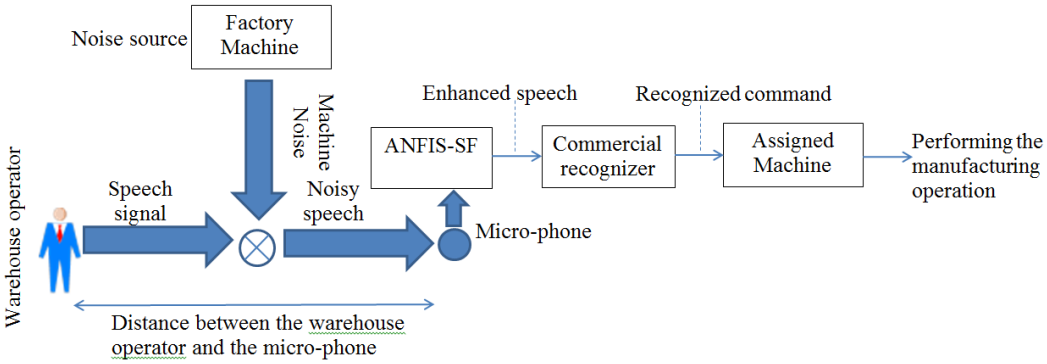

Figure 2 Configuration of the warehouse automation

When the SNR is high, correct commands are likely to be recognized. However, noise always exists in a factory, and inaccurate recognitions are likely to occur. Therefore, ANFIS-SF is used to interface with the commercial recognizer in order to upgrade recognition accuracy. In a factory environment, the SNRs of the noisy speeches received by the commercial recognizer depend on three factors: 1) the volume of the noise produced by the factory machine; 2) the volume of the speech voiced out by the users; and 3) the distance between the user and the commercial recognizer.

The company supporting this research specially developed a Matlab interface for the commercial recognizer whereby the recognized commands can be returned by the Matlab. The noisy environments in a factory were considered, and noise sequences were collected from the Noisex-92 database [33]. If the commercial recognizer is working accurately, factory operators can simply voice out speech commands to control the assigned machine in order to remain focused on the main tasks without requiring both hands to control the machine. Interruptions to their main task can be minimized. Here, the sampling time used for the commercial recognizer is 18 kHz. All the algorithms and computations involved in this study were

implemented using Matlab 2008 in a PC which has a CPU of Intel(R) Core(TM)2 Duo 2.66GHz and a memory of 7.99GB.

*B. **Experimental set-up***

Five speech commands voiced by ten people, eight males and two females, were recorded in a recording studio which is assumed to be noise free. Then, the recordings were contaminated artificially with two types of factory noise, namely machinery noise and engine noise, in order to simulate the noisy environment of factories. The machinery noise was recorded near plate-cutting and electrical welding equipment, and the engine noise was recorded in the engine room while the engine was running. The recorded speech commands were contaminated at different SNRs, so as to simulate the real environment in which the people voice the commands with different volumes and at different distances from the commercial recognizer.

To develop the ANFIS-SF, the number of fuzzy rule, n_{rule}, used in ANFIS-SF is 3. The following parameters, which can be found in reference [47], were implemented in the PSO: the number of particles in the swarm was 100; the number of elements in the particle was 30; both the acceleration constants φ_1 and φ_2 were set at 2.05; the maximum velocity $v_{\max}$ was 0.2; the pre-defined number of generations was 100. Based on the results in [47], these parameters can produce satisfactory results when solving both parameterized and combinatorial problems. Therefore, these parameters are used in this research.

Cross-validations were used to evaluate the effectiveness of the ANFIS-SF, where the cross-validations were repeated five times. The training set was generated by the seven male speech commands and the one female speech command, which were contaminated with various SNRs. The test set was generated by the remaining one male and one female speech commands contaminated with various SNRs. Hence, the performance of the ANFIS-SF can be evaluated in terms of the recognition accuracy under multi-conditions with various SNRs and multi-users. Also, it simulates a practical situation whereby the ANFIS-SF is trained by a group of regular users and is used by new users. It can therefore evaluate whether the trained ANFIS-SF can work accurately with respect to those new users.

After testing the performance of the commercial recognizer by gradually decreasing the SNRs, it was found that the commercial recognizer performs poorly under machinery noise and engine noise with similar SNRs. Therefore, both training sets and test sets were contaminated with similar SNRs for both types of noise. The training set was contaminated with the SNRs of -5dB, -4dB, -3dB, -2dB, -1dB, 0dB, 1dB, 2dB, 3dB, 4dB and 5dB, while the test set was contaminated with the SNRs of -5.5dB, -4.5dB, -3.5dB, -2.5dB, -1.5dB, -0.5dB, 0.5dB, 1.5dB, 2.5dB, 3.5dB, 4.5dB. Therefore, 11 SNR cases were considered for both training and test sets. The SNRs of the test set was smaller than those of the training set with 0.5dB in order to evaluate the recognition accuracy under different SNRs.

In the five cross-validations, there were 5 commands and 8 persons involved in training, so 120 training sequences were involved in each SNR case. While there were 11 SNR cases for training, there were a total of 1320 training sequences involved in training. For testing, there were 5 commands and 2 persons involved, so each SNR case had 50 test sequences. For testing, 11 SNR cases were involved with a total of 550 test sequences.

C. ***Experimental results***

Figure 3 shows the recognition accuracy against machinery noise, where the recognition accuracy is obtained based on the training data for the cross validation. It shows the recognition accuracy obtained solely by the commercial recognizer, ANFIS-SF, and the singly-trained sigmoid filters namely SG-(-5), SG-(0), and SG-(5) which are optimized with respect to a single condition for low SNR (i.e. -5dB), medium SNR (i.e. 0db) and high SNR (i.e. 5dB) respectively. Apart from those, a commonly used noise suppression filter, namely Wiener filter [35], was used for the comparison. The result shows that the commercial recognizer can achieve 65% recognition accuracy when no enhancement method is used under the SNR of 5dB. The performance dropped to about 30% and 16% recognition accuracy when the SNRs decreased to 4dB and -2dB respectively. At -5dB, the recognition accuracy dropped to about 5%. When the Wiener filter is used, recognition accuracy can be improved slightly. However, much better speech recognition can generally be obtained by the sigmoid filters. When working under SNR of -5dB, SG-(-5) performed better than SG-(0),

SG-(5) and Wiener filter. When working under SNR of 0dB and 5dB, SG-(0) and SG-(5) worked better than the others respectively. This result indicates that the sigmoid filters which were developed for single conditions can improve the recognition accuracy only when they are operated under their developed single conditions.

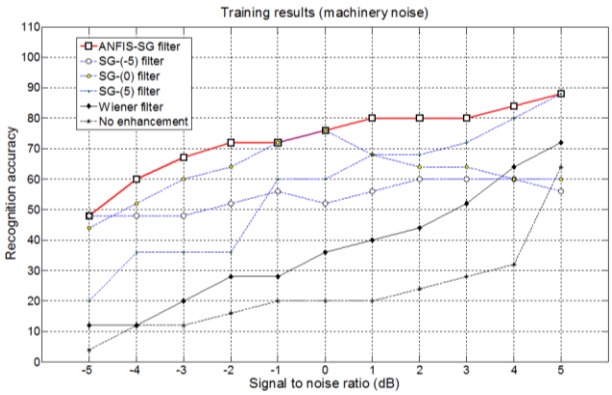

Figure 3 Training results for machinery noise

Figure 3 also shows that the ANFIS-SF can align the outcomes obtained by the best sigmoid filter developed for a single condition. As ANFIS-SF was developed to maximize the recognition accuracy under multi-conditions formulated in (6), the enhanced speech produced by ANFIS-SF can fit well into the commercial recognizer working under multi-conditions. Hence, ANFIS-SF is the best among all filters for improving the recognition accuracy with respect to multi-conditions.

Figure 4 shows the recognition accuracies for the test set which is independent of the training set. Poor speech recognitions were indicated by the commercial recognizer with no improvement. Slightly better results can be obtained by the Wiener filter. The sigmoid filters, SG-(-5), SG-(0) and SG-(5), perform well when working under the SNR conditions in which they were developed. The ANFIS-SF is generally the best sigmoid filter of all the filters used for the test sets. Hence, this result indicates that the commercial recognizer interfaced with ANFIS-SF can work well with the speech commands voiced by the users which have not been involved in the training.

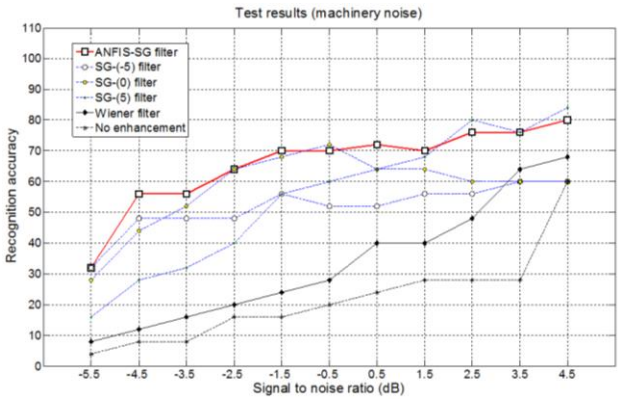

Figure 4 Test results for machinery noise

Figures 5-6 indicate similar results in that the accuracy of the commercial recognizers can be enhanced when the ANFIS-SF were interfaced when working under engine noise with multi-conditions. Results further show that ANFIS-SF outperforms the Wiener filter and the three sigmoid filters, SG-(-5), SG-(0) and SG-(5), which were trained under a single condition. These results further demonstrate that ANFIS-SF tends to satisfy multi-conditions, while SG-(-5), SG-(0) and SG-(5) satisfy only single conditions.

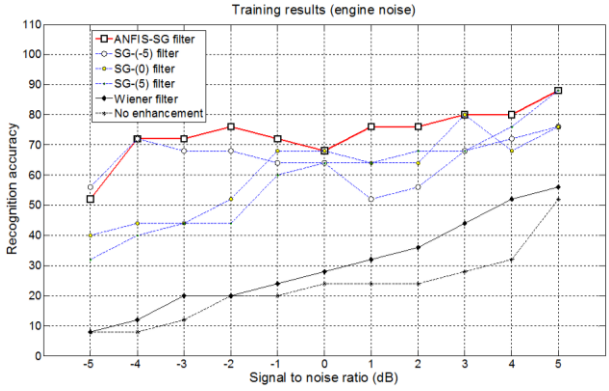

Figure 5 Training results for engine noise

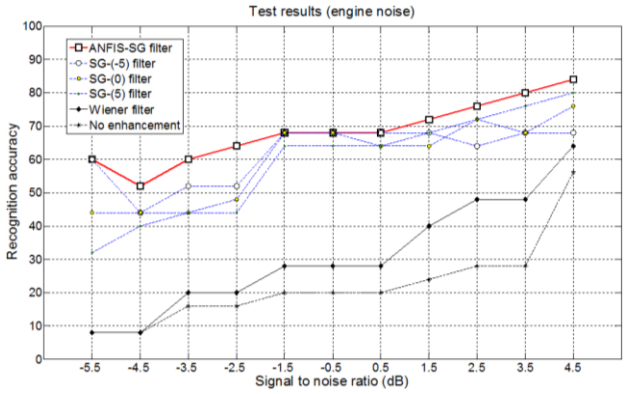

Figure 6 Test results for engine noise

The average computational time required when using solely the commercial recognizer, the ANFIS-SF and the single sigmoid filter for processing a piece of speech command was recorded as 0.0228 seconds, 0.08014 seconds and 0.0809 seconds respectively. It shows that the computational time used by the ANFIS-SF and SG-filter are similar and can be implemented in real time. Hence, it can also be concluded that better recognition accuracy can be obtained by ANFIS-SF which required slightly more computational time.

V. Conclusion and Further Work

In this paper, a hybrid noise suppression filter, namely ANFIS-SF, is proposed based on the mechanisms of sigmoid filter and ANFIS, in order to improve the accuracy of commercial speech recognizers working under multi-noisy conditions. First, a set of optimal filter parameters for the sigmoid filter is determined where each works effectively under a single noisy condition. Then, an ANFIS, is developed based on those optimal filter parameters, in order to perform a functional map between filter parameters and noisy conditions. Based on the functional map, filter parameters are determined for any noisy conditions; hence, ANFIS-SF can work under multi-noisy conditions with varying noisy environments. The effectiveness of the ANFIS-SF is evaluated based on a commercial recognizer. Results show that improvement in terms of speech recognition can be achieved when the ANFIS-SF is interfaced with the commercial recognizer when operating in multi-noisy conditions in factory environments. Also, the computational effort required by ANFIS-SF is

implementable in real-time.

In the future, the mechanism of the proposed ANFIS-SF can be enhanced by conducting the following research:

1) The ANFIS-SF was developed for speech enhancement where only stationary noise can be filtered. It can be further enhanced by integrating with the mechanisms of multi-channel filters, in order to filter non-stationary noise such as a person's conversation, songs or music.
2) The ANFIS-SF will be implemented on an embedded system in order to evaluate the capability of ANFIS-SF in real time, where the filter parameters are adapted in a noisy environment which is time-varying. Based on the embedded system, both the computational time and recognition accuracy can be optimized in real time. Hence, implementation of speech enhancement approaches for speech recognition can move a step forward.
3) Enhanced speech generated by the ANFIS-SF will be analyzed based on well-established acoustic criteria [17]. Based on the findings, the relationship between those acoustic criteria and speech recognition is expected to be developed. Hence, speech recognition can be maximized by optimizing those acoustic criteria which are continuous functions and are easier to optimize than by directly optimizing with respect to the speech recognition.

Acknowledgment

The third and fourth authors were supported by RGC Grant PolyU (5301/12E).

REFERENCES

1. Aberdeen Group, Warehouse automation – What's really working for pallet, case and piece-pick operations, Technical Report, Aberdeen Group, 2007.
2. F. Alonso-Martin and M.A. Salichs, Integration of a voice recognition system in a social robot, Cybernetic and Systems, vol. 42, no. 4 pp. 215-245, 2011.

3. I. Andrianakis and P. White, Speech spectral amplitude estimators using optimally shaped gamma and chi priors, Speech Communication, vol. 51, no. 1, pp. 1-14, 2009.

4. S. Boll, Suppression of acoustic noise in speech using spectral subtraction, IEEE Transaction on Acoustic, Speech, and Signal Processing, 27, pp. 113-120, 1979.

5. F. Bergh and A.P. Engelbrecht, A study of particle swarm optimization particle trajectories, Information Sciences, vol. 176, no. 8, pp.937-971, 2006.

6. C. Breithaupt, M. Krawczyk, R. Martin, Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech, Proceedings on the IEEE International Conference on Acoustic, Speech, and Signal Processing, pp. 4037–4040, 2008.

7. Y. Ephraim and D. Malah, Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator, IEEE Transactions on Acoustic, Speech, Signal Process., vol. 32, no. 6, pp. 1109–1121, 1984.

8. O. Cappe, Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor, IEEE Transaction on Speech and Audio Process, vol. 2, no. 2, pp. 345-349, 1994.

9. K.Y. Chan, K.F.C. Yiu, T.S. Dillon, S. Nordholm and S.H. Ling, Enhancement of speech recognitions for control automation using an intelligent particle swarm optimization, IEEE Transactions on Industrial Informatics, 2012

10. Z. Chen, C. Shan and H. Zhu, Adaptive fuzzy sliding mode control algorithm for a non-affine nonlinear system, IEEE Transactions on Industrial Informatics, vol. 3, no. 4, pp. 302-311, 2007.

11. R.C. Eberhart and Y. Shi, Comparison between genetic algorithms and particle swarm optimization, in Evolutionary Programming VII. New York: Springer-Verlag, LNCS, vol. 1447, pp. 611-616, 1998.

12. R.C. Eberhart, and Y. Shi, Comparing inertia weights and constriction factors in particle swarm optimization, Proc. IEEE Cong. Evolutionary Computing, vol. 1, pp. 84-88, 2000.

13. A. Giaquinto, G. Fornarelli, G. Brunetti and G. Acciani, A neurofuzzy method for the evaluation of soldering global quality index, IEEE Transaction on Industrial Informatics, vol. 51, no. 1, pp. 56-66, 2009.

14. Q.H. He, S. Kwong, K.F. Man, and K.S. Tang, An improved maximum model distance approach for HMM-based speech recognition systems, Pattern Recognition, vol. 33, no. 10, pp. 1749-1758, 2000.
15. J.S. Hu, C.C. Cheng, W.H. Liu and C.H. Yang, A robust adaptive speech enhancement system for vehicular applications, IEEE Transactions on Consumer Electronics, vol. 52, no. 3, pp. 1069-1077, 2006.
16. Y. Hu, P.C. Loizou, N. Li and K. Kasturi, Use of a sigmoidal-shaped function for noise attenuation in cochlear implants, JASA Express Letters, 122(4), pp. 128-134, 2007.
17. Y. Hu and P. Loizou, Evaluation of objective quality measures for speech enhancement, IEEE Transactions on Audio, Speech, and Language Process, vol. 16, no. 1, pp. 229-238, 2008.
18. J.S.R. Jang, "ANFIS: adaptive-network-based fuzzy inference system", IEEE Transactions on Systems, Man and Cybernetics, vol. 23, no. 3, 665- 685, 1993.
19. C.F. Juang, C.T. Chiou and C.L. Lai, Hierarchical singleton type recurrent neural fuzzy networks for noisy speech recognition, IEEE Transactions on Neural Networks, vol. 18, no. 3, pp. 833-843, 2007.
20. S. Kwong, and Qianhua He, The Use of Adaptive Frame for Speech Recognition, EURASIP Journal of Advanced Signal Processing, vol. 2, pp. 82-88, 2001.
21. H.K. Lam and F. Leung, Design and training for combinational neural logic systems, IEEE Transactions on Industrial Electronics, vol. 54, no. 1, pp. 612–619, Feb. 2007.
22. W. Li, D. Wang and T. Chai, Flame image based burning state recognition for sintering process of rotary kiln using heterogeneous features and fuzzy integral, IEEE Transactions on Industrical Informatics, 2012.
23. L. Lu, A. Ghoshal and S. Renals, Regularized subspace Gaussian mixture models for speech recognition, IEEE Signal Processing Letters, vol. 18, no. 7, pp. 419-422, 2011.
24. J. Noyes and A. Starr, Use of automatic speech recognition: current and potential applications, pp. 203-208, 1996.
25. K.E. Parsopoulos and M.N. Vrahatis, On the computation of all global minimizers through particle swarm optimization, IEEE Transactions on Evolutionary Computation, vol. 8, no. 3, pp. 211-224, 2004.

26. J.N. Pires, Robot-by-voice: experiments on commanding an industrial robot using the human voice, An International Journal of Industrial Robot, vol. 32. no. 6, pp. 1159-1320, 2005.

27. Y. Qian, J. Liu and M.T. Johnson, Efficient embedded speech recognition for very large vocabulary Mandarin car-navigation systems, IEEE Transactions on Consumer Electronics, vol. 55, no. 3, 1496-500, 2009.

28. Sensory, INC, Company, Making electronic devices talk and hear, 2010.

29. M. L. Seltzer, B. Raj, and R. M. Stern, Likelihood-maximizing beamforming for robust hands-free speech recognition, IEEE Transactions on Speech and Audio Processing, vol. 12, no. 5, pp. 489–498, 2004.

30. G.S.V.S. Sivaram, S.K. Nemala, N. Mesgarani and H. Hermansky, Data-driven and feedback-based spectro temporal features for speech recognition, IEEE Signal Processing Letter, vol. 17, no. 11, pp. 957-960, 2010.

31. H. Suh and T.W. Kim, Fuzzy membership function based neural networks with applications to visual servoing of robot manipulators, IEEE Transactions on Fuzzy Systems, vol. 2, no. 3, pp. 203-220, 1994.

32. K. Thramboulidis, Model-integrated mechatronics – toward a new paradigm in the development of manufacturing systems, IEEE Transactions on Industrial Informatics, vol. 1, no. 1, pp. 54-61, 2005.

33. 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 Communication, vol. 12, pp. 247–251, 1993.

34. J. Wesley, Fuzzy and Neural Approaches in Engineering, Hines New York, 1997.

35. B.Widrow and S. D. Stearns, Adaptive Signal Processing, Englewood NJ: Prentice-Hall, 1985.

36. F. Yao, Z.Y. Dong, K. Meng, Z. Xu, H.H.C. Iu and K.P. Wong, Quantum inspired particle swarm optimization for power system operations considering wind power uncertainty and carbon tax in Australia, IEEE Transactions on Industrial Informatics, 2012.

37. S.H. Yeung, W.S. Chan, K.T. Ng and K.F. Man, Computational optimization algorithms for antennas and RF/mircrowave circuit design: an overview, IEEE Transactions on Industrial Informatics, vol. 8, no. 2, pp. 216-227, 2012.

38. P.C. Yong, S. Nordholm and H.H. Dam, Noise estimation based on soft decisions and conditional smoothing for speech enhancement, Proceeding on International Workshop on Acoustic Signal Enhancement, pp. 53-56, 2012.

39. P.C. Yong, S. Nordholm and H.H. Dam, Optimization and evaluation of sigmoid function with a prior SNR estimate for real-time speech enhancement, Speech Communication, 2012.

40. J. Zeng and Z.Q. Liu, Yype-2 fuzzy hidden markov models and their Application to speech recognition, IEEE Transactions on Fuzzy Systems, vol. 14, no. 3, pp. 454-467, 2006.

41. M. Zeltzer, B. Raj, and R. Stern, Likelihood-maximizing beamforming for robust hands-free speech recognition, IEEE Transactions on Speech and Audio processing, vol. 12, no. 5, pp. 489–498, 2004.

42. S.X. Zhang, A. Ragni and M.J.F. Gales, Structured log linear models for noise robust speech recognition, IEEE Signal Processing Letters, vol. 17, no. 11, pp. 945-748, 2010.

43. J.H. Zhao, F. Wen, Z.Y. Dong, Y. Xue and K.P. Wong, Optimal dispatch of electric vehicles and wind power using enhanced particle swarm optimization, IEEE Transactions Industrial Informatics, 2012.

44. A. Cohen and D. Graupe, Speech recognition and control system for the severely disabled, Journal of Biomedical Engineering, vol. 2, no. 2, pp. 97-107, 1980.

45. R.I. Damper, Speech control of assistive devices for the physically disabled, IEEE International Conference on Acoustic, Speech and Signal Processing, pp. 653-656, 1986.

46. A.S. AIMejrad, Design of an intelligent system for speech monitoring and treatment of low and excessive vocal intensity.

47. M.O. Neill and A. Brabazon, Grammatical Swarm: The generation of programs by social programming, Natural Computing, vol. 5, pp. 443-462, 2006.

48. C.K. Kwong, T.C. Wong and K.Y. Chan, A methodology of generating customer satisfaction models for new product development using a neuro-fuzzy approach, Expert Systems and Applications, vol. 36, no. 8, pp. 11262-11270, 2009.