Applied Soft Computing 14 (2014) 132–139

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

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

Kit Yan Chan [a,*], Pei Chee Yong [a], Sven Nordholm [a], Cedric K.F. Yiu [b], Hak Keung Lam [c]

[a] Department of Electrical and Computer Engineering, Curtin University, Perth, Australia
[b] Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong
[c] Department of Informatics, King's College London, United Kingdom

ARTICLE INFO

*Article history:*
Received 25 January 2013
Received in revised form 28 May 2013
Accepted 28 May 2013
Available online 13 June 2013

*Keywords:*
Fuzzy neural networks
Noise suppression filter
ANFIS
Speech recognition
Commercial speech recognizer
Sigmoid filter
Speech enhancement

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.



## 1. 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 [14,23], identification mechanisms for capturing significant speech features for recognition [20,30], 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

* Corresponding author. Tel.: +618 9266 2945.
*E-mail address:* chankityan1811@hotmail.com (K.Y. Chan).

1568-4946/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.asoc.2013.05.017

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 [3,6], 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 [10,13,22,31,48]. 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.

## 2. Enhancement of recognition accuracy under multi-conditions

A commercial recognizer, $\Re(\cdot)$, is designed to identify $n$ inbuilt speech commands, $\{u^1, u^2, \ldots, u^n\}$ 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 $\Re(\cdot)$ be denoted as,

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

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; $t = 1, 2, \ldots, m$; and $m$ is the number of samples. Let $\hat{i}$ be the recognized command from $\Re(\cdot)$ for the noisy speech, $x_j^i(t)$, which is given by,

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

where $x_j^i = [x_j^i(1), x_j^i(2), \ldots, x_j^i(m)]$. 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 $\Re(\cdot)$. A sigmoid filter, namely $G_{SIG}(\omega, \ell, \bar{\kappa}(\sigma))$ with three filter parameters, $\bar{\kappa}(\sigma) = [\kappa_1(\sigma), \kappa_2(\sigma), \kappa_3(\sigma)]$, can be used to enhance the accuracy of $\Re(\cdot)$, where $\omega$ is a real angular center frequency given by $\omega \in [\omega_0, \omega_1, \ldots, \omega_{K-1}]$; $\ell$ is the time frame index given by $\ell \in [0, 1, \ldots, L-1]$; $K$ is the number of bands; $L$ is the number of frames; and $\sigma$ is the estimated degrees of signal to noise ratio (SNR).

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

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