ENGINEERING JOURNAL

Article

Level Crossing Control: A Novel Method Using Sound Recognition

Sabur Ajibola Alim*, Nahrul Khair B. Alang Md. Rashid, and Md. Mozasser Rahman

Mechatronics Engineering Department, International Islamic University Malaysia, Kuala Lumpur 50728, Malaysia
*E-mail: moaj1st@yahoo.com (Corresponding author)

Abstract. The level crossing (LX) or railway crossing being an intersection between a public road and a railway line, can be controlled actively or passively. Sound recognition can be used to actively control a level crossing. A system is proposed in this study for the use of sound to control a LX. This proposed system uses Mel Frequency Cepstral Coefficient (MFCC) as feature extractor, and Recurrent Neural Network (RNN) as classifier. The proposed system has shown a great potential that could be harnessed to contribute to the reduction in the loss of lives and properties at the LX.

Keywords: Level crossing, Mel frequency cepstral coefficient (MFCC), recurrent neural network (RNN).

ENGINEERING JOURNAL Volume 17 Issue 3
Received 28 October 2012
Accepted 13 February 2013
Published 1 July 2013
Online at http://www.engj.org/
DOI:10.4186/ej.2013.17.3.113

DOI:10.4186/ej.2013.17.3.113

1. Introduction

A level crossing (LX) is an intersection between a railway line and a public road. LX can be described as either passive or automated based on the protection principle. Passive LX has no protection system whereas the automated one is equipped with at least a set of protection equipment to protect road users from the passing train [1]. Accidents, which include deaths and serious injuries to road users and railway passengers, which occur at level crossings are usually severe. Moreover, these mishaps incur a heavy financial burden on the railway authorities [2] and operators.

Active control of vehicular and pedestrian traffic at a level crossing uses flashing lights, bells, barrier arms, gates or a combination of these devices. Control devices of this nature are triggered manually or automatically by the approaching train. Passive control of vehicular and pedestrian traffic is accomplished by the provision of signs not activated by an approaching train or manually. The purpose is to indicate to the road users to check for the approach of trains prior to crossing the rail lines.

Accordingly, sound recognition is a process of identifying the source or origin of sound and is related to speech recognition. However, unlike speech which has definite structure (vowels, consonants, phonemes) or music which has harmonic structure (notes, rhythm, timbre) [3], sound is typically made up of a mixture of other sounds, that are characterized by spectral peaks, such as insect chirpings. Sounds are unstructured and comparable to noise, variably composed and thus, models are difficult to build for them [4]. Similarly, sounds are known for their randomness and high variance [5].

2. Operation of the Proposed System

A very sensitive unidirectional microphone would be used to acquire the sound at a predetermined distance from the level crossing. The distance equal to the length of the train is preferred as this would make it easy to apply the brakes in case of any problem. The sound recognition procedures are followed sequentially. Starting from pre-emphasis filtering and stopping with classification and identification. If the sound of the train is identified, then the level crossing protection sequence is activated. As soon as the sound of train is identified, the traffic lights are turned on, starting with yellow light and subsequently the red light. Simultaneously as the traffic lights are activated, warning bells are sounded for the blind road users. Immediately the level crossing is at safety, the train passes through. Since this system is independent of the train control system, a set of cameras would be installed at the level crossing. This is to enable the train driver and the train management center to confirm the status of the level crossing and avoid accident.

3. Methodology

3.1. Sample Collection

At the level crossing, several sounds are made simultaneously. As a result, there is the need to mix up some of the individual sounds that are heard around the level crossing to match up with real life scenario. The individual sounds (aircraft, car, rain, train, thunder) were obtained from both online databases and live recording. During the sound data collection, the sounds were sampled at 11025Hz and the Nyquist sampling theorem was fulfilled. When the sounds were mixed, since the sound of the train is the needed sound, the other sound is taken to be noise (unwanted sound). This is similar to the level crossing which is the intended area of application, there, the sound of the train will be required and all other sounds would be regarded as noise. The sound mixing was done using the NCH software's "WavePad Sound Editor Masters Edition v 5.10". The sound mixtures were saved in an uncompressed .wav format, sampled at 11025 Hz and 16 bit. The sounds were mixed in a manner in which they overlap over each other i.e. the sound mixture plays as a single sound entity.

The sound of the train was mixed with some of the sounds that are commonly found around the level, they are sounds of the aircraft, car, rain and thunder. The two sound mixtures are as follow, train+aircraft, train+car, train+rain and train+thunder. After the result from the mixtures of two sounds was achieved, it was decided to make a mixture of three sounds. This three sound mixture moves the proposed system much closer to real life scenario. The result from this test will help to determine how this system can fit in either as a separate level crossing assembly that is reliable or as a system incorporated into one of the

existing systems as a backup system and to be able to recommend areas of this work that need further research. The sounds were mixed as follows: train+aircraft+car, train+car+rain and train+rain+thunder.

3.2. Mel Frequency Cepstral Coefficients (MFCC)

Mel Frequency Cepstral Coefficients (MFCC) was initially proposed for speech recognition to identify monosyllabic words in continuously spoken sentences but not for Speaker Identification. MFCCs calculation is based on the human auditory system aiming for artificial implementation of the ear's physiology assuming that the human ear can be a good speaker recognizer [6]. Its features are based on the known frequency variations in the human ear's critical bandwidths. Filters linearly spaced at low frequency and logarithmically at high frequency are being used to capture the phonetically essential characteristics of speech [7]. A speech signal usually consist of tones with dissimilar frequencies, each tone having an actual frequency, f (Hz) and a subjective pitch measured on the mel scale. The mel frequency scale is a linear frequency spacing lower than 1000Hz and a logarithmic spacing higher than 1000 Hz. As a reference point, a pitch of 1 kHz tone, 40 dB more than the perceptual hearing threshold, defines 1000 mels [8].

3.3. Recurrent Neural Network (RNN)

The architecture of Recurrent Neural Network requires a number of input layer neurons, hidden layer neurons and output layer neuron. Also, it has a feedback loop with a single delay in the hidden layer. The output of the input layer is fed back into it as part of its inputs. The Layer-Recurrent Network (LRN) used is a simplified version of the Elman network. This network is a dynamic network having memory and can be trained to learn sequential or time-varying patterns. For the 2 sounds mixture, there were 300 input neurons to accommodate the input feature vectors with each containing 12 coefficients per frame for five frames. The number of nodes in hidden layer was varied to get best result, which then was selected; the number of output nodes is 4 nodes. In case of the 3 sounds mixture, there were 225 input neurons to accommodate the input feature vectors with each containing 12 coefficients per frame for five frames. The number of nodes in hidden layer was varied to get best result, which was then selected; the number of output nodes is 3 nodes.

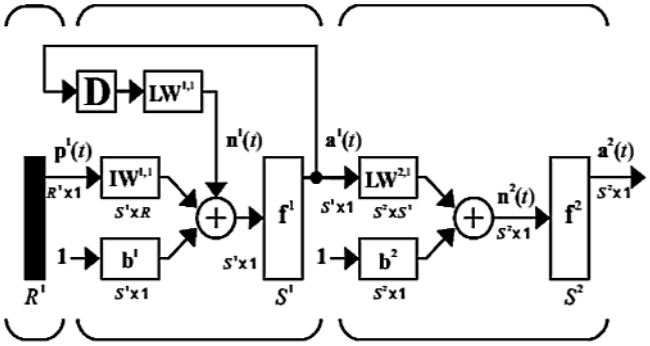

Fig. 1. Architecture of Recurrent Neural Network used in the research; D- delay, p'(t)- time varying input, I- net function of input layer, L- net function of hidden layer, b- bias, a- layer output.

As of the 2 sounds mixture, the sound of train has to be used to test the network using 100 samples. Therefore, from the testing of the sound mixtures, the confusion matrix was plotted. For the 3 sounds mixture, the sound train has to be used to test the network using 75 samples. Therefore, from the testing of the sound mixtures, the confusion matrix was plotted. The confusion matrix plot was then used for the evaluation of the sensitivity of each sound mixture to the sound of train as input and the rate at which each sound mixture misclassifies the sound of train.

3.4. Training Algorithm

The backpropagation algorithm was used in the training of the network. The computation was made faster with the selection faster variant of the conventional backpropagation, namely the scaled conjugate gradient

algorithm (trainscg). This algorithm uses standard numerical optimization techniques. The scaled conjugate gradient algorithm (SCG) was designed to reduce the time-consumed in line search and combines both the model-trust region approach (Levenberg-Marquardt algorithm) and the conjugate gradient approach [9].

SCG is a second order conjugate gradient algorithm that helps minimize goal functions of several variables. It uses a step size scaling mechanism that avoids a time consuming line-search per learning iteration, which makes the algorithm faster than other second order algorithms. SCG shows super linear convergence on most problems [10]. SCG does not include any user dependent parameters which values are crucial for the success of the algorithm. This is a major advantage compared to the line search based algorithms which include those kinds of parameters. The SCG algorithm is as shown below [9]:

1. Choose weight vector $\grave{w}_1$ and scalars $0 < \sigma \leq 10^{-4}$, $0 < \lambda_1 \leq 10^{-6}$, $\bar{\lambda}_1$
   Set $\tilde{p}_1 = \tilde{r}_1 = -E'(\widetilde{w}_1)$, k=1 and success = true
2. If success =true, the calculate second order information:
   $\sigma_k = \sigma/|\tilde{p}_k|,$ (1)
   $\tilde{s}_k = \big(E'(\widetilde{w}_k + \sigma_k\tilde{p}_k) - E'(\widetilde{w}_k)\big)/\sigma_k,$ (2)
   $\delta_k = \tilde{p}_k^T\tilde{s}_k.$ (3)
3. Scale δ_k: $\delta_k = \delta_k + \big(\lambda_k - \bar{\lambda}_k\big)|\tilde{p}_k|^2$ (4)
4. If $\delta_k \leq 0$ then make the Hessian matrix positive definite:
   $\bar{\lambda}_1 = 2\big((\lambda_k - \delta_k)|\tilde{p}_k|^2\big)$ (5)
   $\delta_k = -\delta_k + \lambda_k|\tilde{p}_k|^2$ (6)
   $\lambda_k = \bar{\lambda}_k$ (7)
5. Calculate step size:
   $\mu_k = \tilde{p}_k^T\tilde{r}_k,$ (8)
   $\alpha_k = \mu_k/\delta_k$ (9)
6. Calculate the comparison parameter:
   $\Delta_k = 2\delta_k\,[E(\widetilde{w}_k) - E(\widetilde{w}_k + \alpha_k\tilde{p}_k)]/\mu_k^2$ (10)
7. If $\Delta_k \geq 0$, then a successful reduction in error can be made:
   $\widetilde{w}_{k+1} = \widetilde{w}_k + \alpha_k\tilde{p}_k,$
   $\tilde{r}_{k+1} = -E'(\widetilde{w}_{k+1}),$
   $\bar{\lambda}_k = 0$, then success = true
   If k mod N = 0 then restart algorithm:
   $\tilde{p}_{k+1} = \tilde{r}_{k+1}$
   else:
   $\beta_k = \big(|\tilde{r}_{k+1}|^2 - \tilde{r}_{k+1}^T\tilde{r}_k\big)/\mu_k,$
   $\tilde{p}_{k+1} = \tilde{r}_{k+1} + \beta_k\tilde{p}_k$
   If $\Delta_k \geq 0.75$, then reduce the scale parameter:
   $\lambda_k = \frac{1}{4}\lambda_k$
   else
   $\bar{\lambda}_k = \lambda_k$
   Success = false.
8. If $\Delta_k < 0.25$, then increase the scale parameter:
   $\lambda_k = \lambda_k + (\delta_k(1 - \Delta_k)/|\tilde{p}_k|^2)$ (11)
9. If the steepest descent direction $\tilde{r}_k \neq 0$, then set k = K+1 and go to 2 else terminate and return to $\widetilde{w}_{k+1}$ as the desired minimum.

σ, λ_1 and $\bar{\lambda}_1$- weight scalar, $-E'(\widetilde{w}_1)$- negative gradient, k- iteration, σ_k- directional gradient, $\tilde{p}_k$- search direction, $\tilde{s}_k$- directional curvature, $\widetilde{w}_k$- weight vector, δ_k- working curvature, λ_k- Lagrange multiplier, α_k- step size, $\tilde{p}_k^T$- transpose of search direction, $\tilde{r}_k$- steepest descent direction, Δ_k- comparison parameter, $\widetilde{w}_{k+1}$- updated weight vector, $\tilde{r}_{k+1}$- updated steepest descent direction and $\tilde{p}_{k+1}$- updated search direction.

4. Results

4.1. Two Sounds Mixture

The performance metrics of the sound of train+aircraft, train+car, train+rain, and train+thunder, using Recurrent Neural Network (RNN) as classifier, with Mel Frequency Cepstral Coefficient (MFCC) as feature extractor are shown in Table 1. The testing was conducted to determine the sound recognition and classifier capability with 40 hidden neurons. The sound of train+rain had the highest sensitivity (86.7%), while the sound of train+car had the lowest sensitivity (56.7%). Similarly, the highest accuracy (90%) was for the sound of train+rain and the lowest accuracy (80.8%) was for the sound of train+car. In addition, the sound of train+car gave the highest (19.2%) misclassification rate, while the sound of train+rain gave the lowest (10%) misclassification rate.

Table 1. Performance of Mel frequency cepstral coefficient (MFCC) using recurrent neural network (RNN) as classifier.

	Sensitivity (%)	Accuracy (%)	Misclassification rate (%)
Train + Aircraft	76.7	84.2	15.8
Train + Car	56.7	80.8	19.2
Train + Rain	86.7	90.0	10.0
Train + Thunder	60.0	85.0	15.0

The misclassification rates for the sounds samples are below 0.2 (20%), therefore they are categorized as negligible. Likewise, the sensitivity of train+car and train+thunder sounds fall between the range of 0.4-0.6 (40-60%), therefore they are grouped as moderate. Furthermore, the sensitivity for the sound of train+aircraft is between the range 0.6 and 0.8 (60-80%), and as result is classified as substantial while the sensitivity for the sound of train+rain falls in range 0.8-1.0 (80-100%) therefore it is classified high [11].

4.2. Three Sounds Mixture

The performance metrics of the sound of train+aircraft+car, train+car+rain and train+rain+thunder, using Recurrent Neural Network (RNN) as classifier, with Mel Frequency Cepstral Coefficient (MFCC) feature extractor is shown in Table 2. The testing was conducted to determine the sound recognition and classifier capability with 40 hidden neurons. The sound of train+aircraft+car gave the highest (90%) sensitivity and the sound of train+car+rain gave the lowest (60%) sensitivity. In addition, the accuracy for the sounds of train+aircraft+car and train+rain+thunder gave the highest (90%) and lowest (80%) respectively. Finally, the sound of train+rain+thunder and train+aircraft+car gave the highest (20%) and lowest (10%) misclassification rates respectively.

Table 2. Performance of Mel frequency cepstral coefficient (MFCC) using recurrent neural network (RNN) as classifier.

	Sensitivity (%)	Accuracy (%)	Misclassification rate (%)
Train + Aircraft + Car	90.0	90.0	10.0
Train + Car + Rain	60.0	81.1	18.9
Train + Rain + Thunder	76.7	80.0	20.0

The misclassification rate of all samples are below 0.2 (20%), therefore they are categorized as negligible. The sensitivity of train+aircraft+car (90%) falls between 0.8 and 1.0 (80-100%), therefore it is denoted as high. Furthermore, the sensitivity for the sound of train+car+rain (60%) is in between 0.4 and 0.6 (40-60%) therefore it is categorized moderate. In addition, the sensitivity for the sound of train+rain+thunder (76.7%) falls between 0.6-0.8 (60-80%), hence, it is classified as substantial [11].

In pattern recognition, it is desirable to have sensitivities that fall in the range 0.8-1.0 (80-100%) that can therefore be categorized as high, while it is also necessary to have misclassification rates to fall in the range 0.0-0.2 (0-20%), it can therefore be classified as negligible. This is because it is very paramount to

design systems that have high sensitivity and low misclassification rate, so that when it is implemented in real life, there is high confidence in the design. However, since MFCC was designed to model the human auditory system, it may have some difficulty in recognizing sounds efficiently. This is because sound is variably composed, similar to noise and has a lot of frequency irregularities as compared to speech or music.

5. Conclusion

The combination of Mel Frequency Cepstral Coefficient (MFCC) with Recurrent Neural Network has given a set of results that are fairly good for both 2 sounds and 3 sounds mixtures. All the sensitivities obtained are in range of moderate to high, while all the misclassification rates are negligible. Consequently, the proposed system, using the feature extractor and classifier combination, has shown a great potential and is recommended for further investigation to ascertain more facts and to standardize its essential parameters.

References

[1] W. Zheng, J. R. Mueller, R. Slovak, and E. Schnieder, "Function modelling and risk analysis of automated level crossing based on national statistical data," in *Proceedings of International Asia Conference on Informatics in Control, Automation and Robotics (CAR)*, Wuhan, China, 2010, pp. 281-284.

[2] S. Z. Ishak, W. L. Yue, and S. Somenahalli, "Level crossing modelling using petri nets approach and π–tool," *Asian Transport Studies*, vol. 1, no. 2, pp. 107-121, 2010.

[3] O. Lartillot and P. Toiviainen, "A matlab toolbox for musical feature extraction from audio," in *Proceedings of International Conference on Digital Audio Effects (DAFx-07)*, Bordeaux, France, 2007, pp. 1-8.

[4] J. R. Rabuñal and J. Dorado, *Artificial neural networks in real-life applications*. Idea group publishing, 2006, pp. 1-375.

[5] S. Chu, S. Narayanan, and J. C. C. Kuo, "Environmental sound recognition using MP-based features," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008)*, Las Vegas, NV, 2008, pp. 1 – 4.

[6] S. Chakroborty, A. Roy, and G. Saha, "Fusion of a complementary feature set with MFCC for improved closed set text-independent speaker identification," in *Proceedings of IEEE International Conference on Industrial Technology (ICIT 2006)*, 2006, pp. 387–390.

[7] J. Lara, "A method of automatic speaker recognition using cepstral features and vectorial quantization," in *Lecture Notes in Computer Science (LNCS)*, vol. 3773, M. Lazo and A. Sanfeliu, Eds. Berlin Heidelberg: Springer-Verlag, 2005, pp. 146-153.

[8] P. P. Kumar, K. S. N. Vardhan, and K. S. R. Krishna, "Performance evaluation of MLP for speech recognition in noisy environments using MFCC & wavelets," *International Journal of Computer Science & Communication (IJCSC)*, vol. 1, no. 2, pp. 41-45, 2010.

[9] M. F. MØller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Networks*, vol. 6, no. 4, pp. 525-533, 1993.

[10] Z. Zakaria, A. Mat Isa, and S. A. A. Suandi, "Study on neural network training algorithm for multiface detection in static images," in *Proceedings of International Conference on Computer, Electrical, and Systems Science, and Engineering (ICCESSE 2010)*, World Academy of Science, Engineering and Technology, Penang, Malaysia, 2010.

[11] J. W. Best, *Research in Education*. Englewood Cliffs, NJ: Prentice-Hall, 1981.