2009 Third International Conference on Network and System Security

Speech recognition enhancement using beamforming and a genetic algorithm

K.Y. Chan
Digital Ecosystems and Business Intelligence Institute, Curtin University of Technology, Perth, Australia
Email: Kit.Chan@curtin.edu.au

K.F.C. Yiu
Department of Applied Mathematics,
The Hong Kong Polytechnic University, Kowloon, Hong Kong, PR China

S.Y. Low, S. Nordholm
Western Australian Telecommunications Research Institute
A joint venture between The University of Western Australia and The University of Curtin, Perth, Australia

S.H. Ling
Centre for Health Technologies
Faculty of Engineering and Information Technology
University of Technology Sydney, NSW
Australia

Abstract— **This paper proposes a genetic algorithm (GA) based beamformer to optimize speech recognition accuracy for a pre-trained speech recognizer. The proposed beamformer is designed to tackle the non-differentiable and non-linear natures of speech recognition by employing the GA algorithm to search for the optimal beamformer weights. Specifically, a population of beamformer weights is reproduced by crossover and mutation until the optimal beamformer weights are obtained. Results show that the speech recognition accuracies can be greatly improved even in noisy environments.**

Keywords- Speech recognition, beamforming, signal enhancement, genetic algorithm.

I. INTRODUCTION

Speech recognition system is widely used in automatic remote control such as logistics warehouse [1], intelligence home [2], banking systems [3] and interactive children books [4]. It is usually based on the hidden Markov model (HMM) chain and is trained to recognize the commands using a large database of speech signals [5]. However, noise contamination is inevitable in real world environment which severely degrades the performance of speech recognition systems [6, 7]. This is especially the case in distant-talking microphones where the signal of interest is usually under the influence of ambient noise and reverberation [8], which causes a mismatch in the recognition process.

One common method to improve the recognition rate is to pre-process the signal via array processing methods such as beamforming [9, 10]. As such, the distortion mentioned above can be compensated spatially by focusing the array to point towards the signal of interest. An example is the delay and sum beamformer where the received microphone signals are time aligned to focus onto the direction of interest [11]. An alternative to the delay and sum beamformer is the adaptive beamformer where the array parameters are optimized in a continuous fashion to minimize the power from all directions, whilst maintaining a unity from the direction of interest [9, 10].

Generally speaking, beamformer design can be viewed as a multi-criteria decision problem, where the two most common criteria are target signal distortion and noise suppression [12]. As pointed out in [13], these two criteria must be carefully optimized to produce a good recognition rate due to the complexity of the recognizer. For instance, an optimum noise suppression capability may not necessarily translate to an increase in the recognition rate as the target signal may be distorted. Likewise, a strict constraint on the target signal distortion may limit the noise suppression capability.

As opposed to optimizing the beamformer in a front-end processor fashion, this paper proposes to optimize the beamformer weights with respect to speech recognition accuracy. Gradient descent methods are the most commonly used technique in performing optimization in which a cost function is well defined. However, the optimization of speech recognition accuracy is a difficult problem since it is discontinuous and highly nonlinear. Also, it is not possible to formulate the problem into an integer programming problem and thus cannot be handled by the gradient descent methods. Genetic algorithm on the other hand offers a promising solution in solving discontinuous and highly nonlinear problems [14, 15]. In this paper, we propose a genetic algorithm to optimize the beamforming weights. Validations were carried out by designing the beamformers for maximizing speech recognition accuracies on two sets of contaminated command signals with poor signal to noise ratios. Results show that the speech recognition accuracy can be improved from low levels to 100% accurate rates by the proposed genetic algorithm.

978-0-7695-3838-9/09 $26.00 © 2009 IEEE
DOI 10.1109/NSS.2009.44

II. THE SIGNAL MODEL

In a typical environment of using a pre-trained speech recognizer based on the principle of hidden Markov model, assume there are a fixed set of n voice commands, denoted by $\{s^1, s^2, \ldots, s^n\}$, built into the dialog between the system and users. A dialog is defined as a finite state machine, which consists of states and transitions. A dialog state represents one conversational interchange between the system and user, typically consisting of a prompt and then the user's response. The system constantly listens to the trigger phrase in the system standby phase. As soon as the user says the general-purpose trigger phrase, the system will respond with an acknowledge tone. The caller is response to specify the desired transaction. The caller responds in variety of ways but must include one of several keywords that define a supported transaction. In the case of a user profile transaction, the application will retrieve the pre-programmed setting of the specified user, and prompt the user with confirmation before going back to the system standby state.

Due to the presence of acoustic noise in the environment, the input commands are usually distorted by a noise, which is given by

$$x^i(k) = s^i(k) + v^i(k), \qquad i = 1,2,\ldots,n, \tag{1}$$

where $v^i(k)$ is the noise signal. Note that the noise signal could include a sum of fixed point noise sources together with a mixture of coherent and incoherent noise sources. A beamformer with M elements in the microphone array is proposed to filter the contaminated signal in which the k-th sample of the signal received by the j-th microphone is represented by:

$$x_j^i(k) = s_j^i(k) + v_j^i(k), \qquad j=1,2,\ldots M, \tag{2}$$

where $s_j^i(k)$ and $v_j^i(k)$ is the k-th sample of the source signal and the noise signal received by the j-th microphone respectively. The output of the beamformer is given by:

$$y^i(k) = \sum_{j=1}^{M} \sum_{l=0}^{L-1} w_j(l) x_j^i(k-l) \tag{3}$$

where L is the filter length of the beamformer. The beamformer matrix $\mathbf{w}$ is represented by:

$$\mathbf{w} = \begin{pmatrix} w_1^T & w_2^T & \ldots & w_M^T \end{pmatrix}^T \tag{4}$$

where $w_j^T = \begin{pmatrix} w_j(0) & w_j(1) & \ldots & w_j(L-1) \end{pmatrix}$, j=1,2,..., M. For the received i-th command, a vector of scores is calculated, denoted by

$$\{L_1(y^i), \cdots, L_n(y^i)\} \tag{5}$$

where $L_j(y_i)$ stands for the likelihood that the received command is the j-th command. With filtering, the estimated command is taken to be

$$\hat{i} = \arg\max_j \{L_j(y^i)\}. \tag{6}$$

Define

$$f_i = \begin{cases} 0 & \text{if } \hat{i} \neq i \\ 1 & \text{if } \hat{i} = i \end{cases},$$

the score of correct recognition for a pre-recorded command set or a calibrated command set recorded in a quiet environment can be calculated as

$$f(\mathbf{w}) = \sum_{i=1}^{n} f_i, \tag{7}$$

where f is clearly a highly nonlinear function of $\mathbf{w}$. To enhance the accuracy of the recognition, it is sufficient to optimizing on the beamformer matrix $\mathbf{w}$ to maximize f. Since $f(\mathbf{w})$ is a non-differentiable and a nonlinear function, gradient-based approaches cannot be used adequately. A genetic algorithm based method, which is proposed to solve the optimization problem (7), is presented in the following section.

III. GENETIC ALGORITHM BASED METHOD

In order to simulate the situation of typical voice control devices, a configuration of four element square microphone array with 30cm apart horizontally and vertically is used and is illustrated in Figure 1. The command source is standing 1m away from the microphone array as illustrated in Figure 2. The near-field noise is placed 1m in front of the array and 1m to the left of the speaker. Each command is super-imposed by the near-field noise and test for correctness by feeding into the speech recognizer. This configuration simulates a real home environment, where a recognizer is placed in front of a user and a noise source like a radio is placed beside the user.

In this test, three various kinds of near-field noises (music noise, radio noise and song noise illustrated in Figure 3, 4 and 5) are used. All the near-field noises and the calibration source signals are recorded in an anechoic environment with a sampling rate of 16kHz. Two sets of commands are created to test the proposed method. The first set consists of names of Christmas songs (*jingle bells*; *santa claus is coming totown*; *sleigh ride*; *let it snow*; *winter wonderland*) typically used in a music-box. This is a typical command set with phrases. We denote this set of commands by Musicbox and the first command, *jingle bells*, is illustrated in Figure 6. The second set of commands is a set of nine single word-based commands using the simple numbers. We denote this set of commands by Numbers, and the first command, *one*, is illustrated in Figure 7.

The actual signal-to-noise ratios are measured by a sound pressure level (SPL) meter. The level of noise is increased gradually until the total recognition accuracy is below 50%. This corresponds to 8dB for the command set Numbers and 2dB for the command set Musicbox. The command set Numbers is based on a single word while the command set Musicbox is based on a phrase consisting of a few words. The spectral patterns of phrases are easier to be recognized

than the ones of the single words. Therefore the command set Musicbox is easier to be recognized than the one with Numbers, and the corresponding signal to noise ratio for the command set Numbers is higher than the command set Musicbox.

For the command set Musicbox, Table 1 shows that the recognition accuracy has fallen below 40% without any enhancement. However, by using the proposed beamforming method, a fairly uniform improvement to 100% can be achieved for almost all the tested noise, which the filter length L=16 is used in all kinds of noises.

Table 1 Correct recognition rates for the command set Musicbox

Noise type	Pure recognition	Recognition with beamformer	Filter length of beamformer (L)
Music	40%	100%	16
Radio	40%	100%	16
Song	40%	100%	16

For the command set Numbers, Table 2 shows that the findings are generally similar to the results for the Musicbox. Clearly the improvement is significant over the recognition without any enhancement. In general, this is not a recommended command set for recognition due to the similarity among commands and the short durations which make the recognitions very difficult. Therefore the longer filtering length L=32 is required to achieve 100% accuracy in radio noise. Nevertheless, the proposed beamforming method still achieves reasonable improvement for this difficult command set. The results demonstrate that the proposed method works well in a real home environment to enhance recognition accuracy.

Table 2 Correct recognition rates for the command set Numbers

Noise type	Pure recognition	Recognition with beamformer	Filter length of beamformer (L)
Music	66.67%	100%	16
Radio	11.11%	100%	32
Song	44.44%	100%	16

IV. CONCLUSION AND FURTHER WORK

A new speech enhancement method, which uses GA to optimize the performance of beamformer, has been proposed to directly maximize speech recognition accuracy. It compensates the deficiencies of the existing enhancement methods, which is designed for minimizing signal distortion or maximizing noise suppression, but cannot directly deal with speech recognition accuracies. The performance of the proposed method is evaluated by using a pre-trained recognizer embedded with two sets of speech commands. Results show that full scored accuracy rates of speech recognitions, which are contaminated by three types of noise, can be achieved. It demonstrated that the performance of speech recognizer can be upgraded by the proposed method while it works on noisy environments as in real situation.

In the future, we will apply our developed GA [16] on solving this speech recognition problem. It is expected that the computational time can be shorter.

Acknowledgement

The authors would like to thank you Digital Ecosystems and Business Intelligence Institute, Curtin University of Technology on supporting this project.

REFERENCES

[1] J. L. Gauvain, J.J. Gangolf, and L. Lamel, Speech recognition for an information kiosk, International Conference on Spoken Language, vol. 2, pp. 849–852, 1996.

[2] A. Burstein, A. Stolzle, and R.W. Brodersen, Using speech recognition in a personal communications system, IEEE International Conference on Communications, vol. 3, pp. 1717–1721, 1992.

[3] T. Isobe, M. Morishima, F. Yoshitani, N. Koizumi and K. Murakami, Voice-activated home banking system and its field trial, International Conference on Spoken Language, vol. 3, pp. 1688–1691, 1996.

[4] A. Hagen, B. Pellom and R. Cole, Children's speech recognition with application to interactive books and tutors, IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 186–191, 2003.

[5] P. Woodland, Speech recognition, IEE Colloquium on Speech and Language Engineering, pp. 1–5, 1998.

[6] D. Giuliani, M. Matassoni, M. Omologo, and P. Svaizer, Hands-free continuous speech recognition in noisy environment using a four microphone array, International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 860–863, 1995.

[7] J. Sirigos, N. Fakotakis, and G. Kokkinakis, Improving environmental robustness of speech recognition using neural networks, International Conference on Digital Signal Processing vol. 2, pp. 575–578, 1997.

[8] 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.

[9] B. D. Van Veen and K. M. Buckley, Beamforming: A versatile approach to spatial filtering, IEEE Acoust., Speech and Signal Process. Magazine, vol. 5, pp. 4–24, 1988.

[10] D. H Johnson. and D. E. Dudgeon, Array Signal Processing: Concepts and Applications. Englewood Cliffs, New Jersey: Prentice-Hall, 1993.

[11] W. Kellermann, A self-steering digital microphone array, IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 5, pp. 3581–3584, 1991.

[12] K. F. C. Yiu, N. Grbi´c, K.L. Teo, and S. Nordholm, A new design method for broadband microphone arrays for speech input in automobiles, IEEE Signal Processing Letters, vol. 9, no. 7, pp. 222–224, 2002.

[13] K. F. C. Yiu., S.Y. Low, K.Y. Chan and S. Nordholm, A multi-filter system for speech enhancement under low signal-to-noise ratios, Journal of Industrial and Management Optimization, 2009.

[14] T. Back, U. Hammel, and H.P. Schwefel, Evolutionary computation: comments on history and current state, IEEE Transactions on Evolutionary Computation, vol. 1, no. 1, pp. 3-17, 1997.

[15] K.F. Man, K.S. Tang and S. Kwong, Genetic algorithms: concepts and applications, IEEE Transactions on Industrial Electronics, vol. 43, no. 5, pp. 519-536, 1996.

[16] K.Y. Chan, K.W. Chan, G.T.Y. Pong, M.E. Aydin, T.C. Fogarty and S.H. Ling, A statistics-based genetic algorithm for quality improvements of power supplies, European Journal of Industrial Engineering, 2009.

Appendix

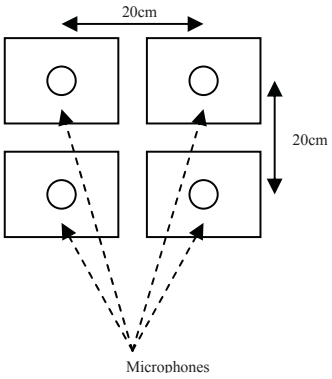

Figure 1 Front view of the microphone array.

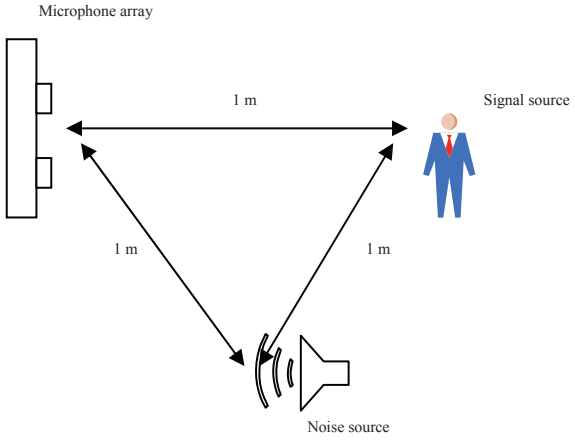

Figure 2 Configuration 1 simulated the environment of home intelligent.

.

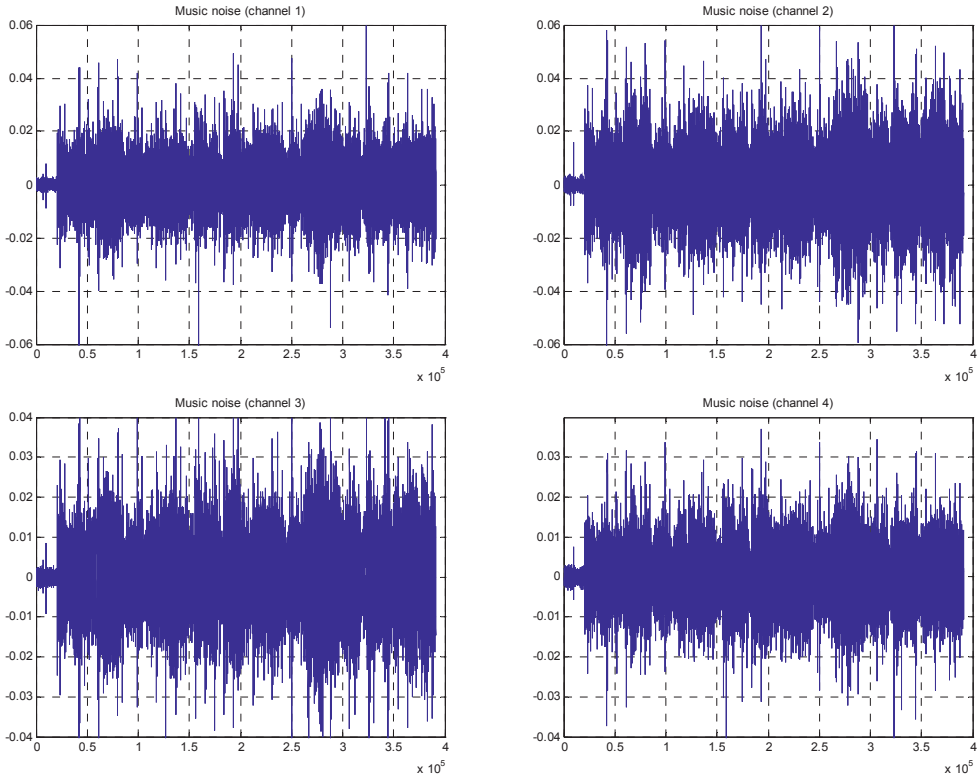

Figure 3 Music noise in the four channels.

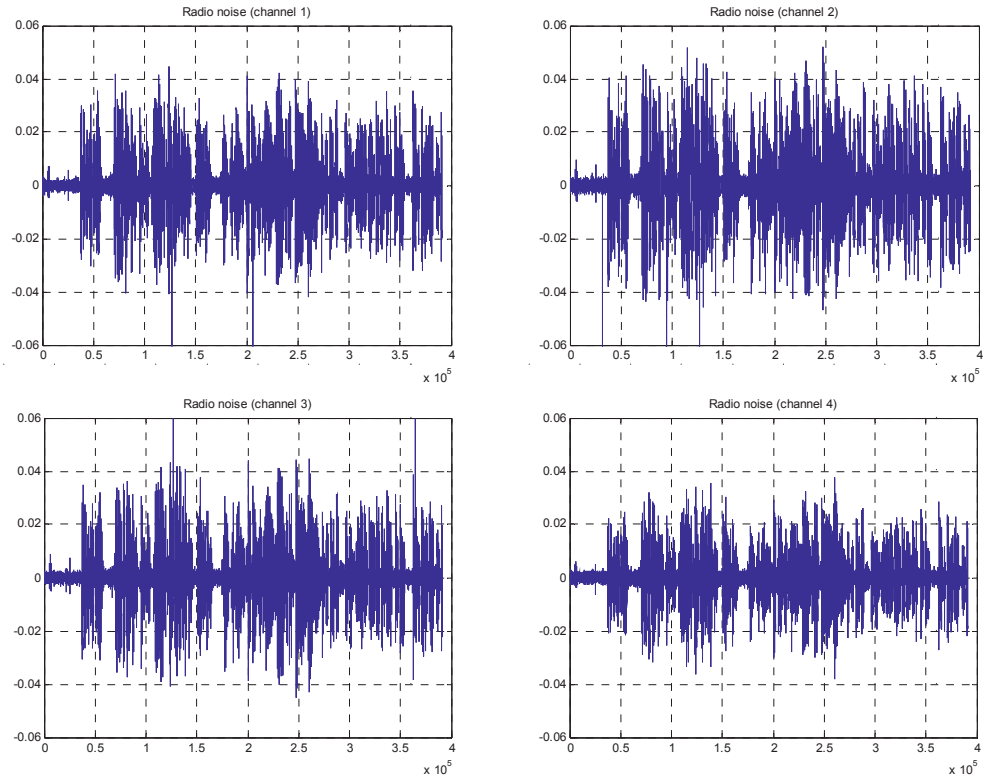

Figure 4 Radio noise in the four channels.

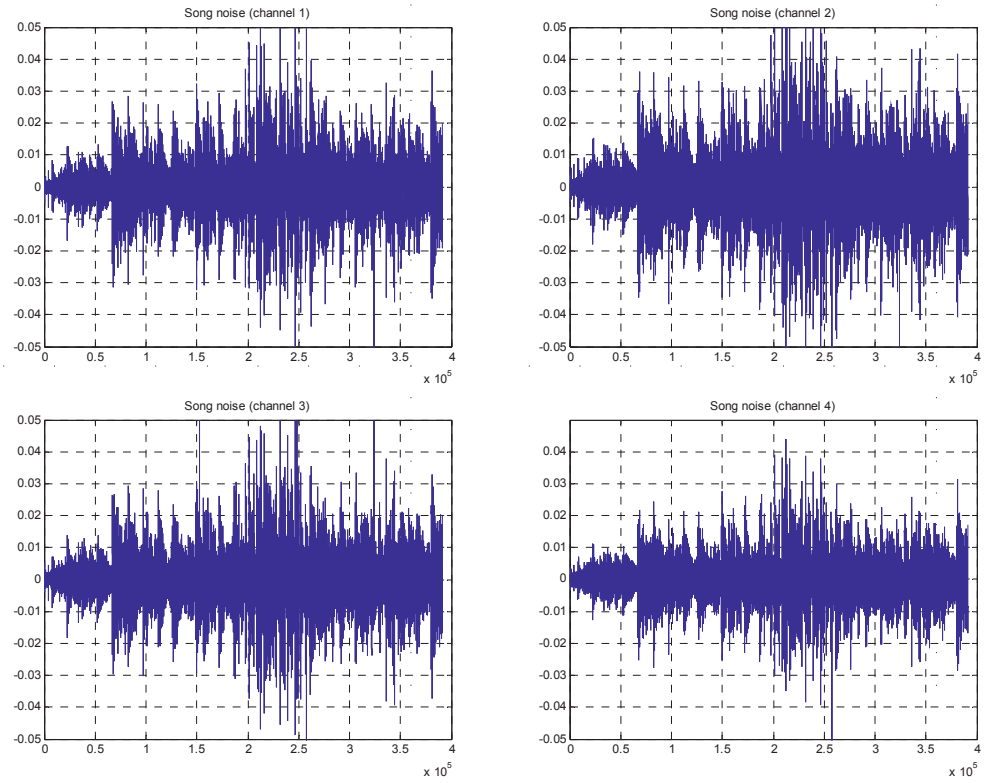

Figure 5 Song noise in the four channels.

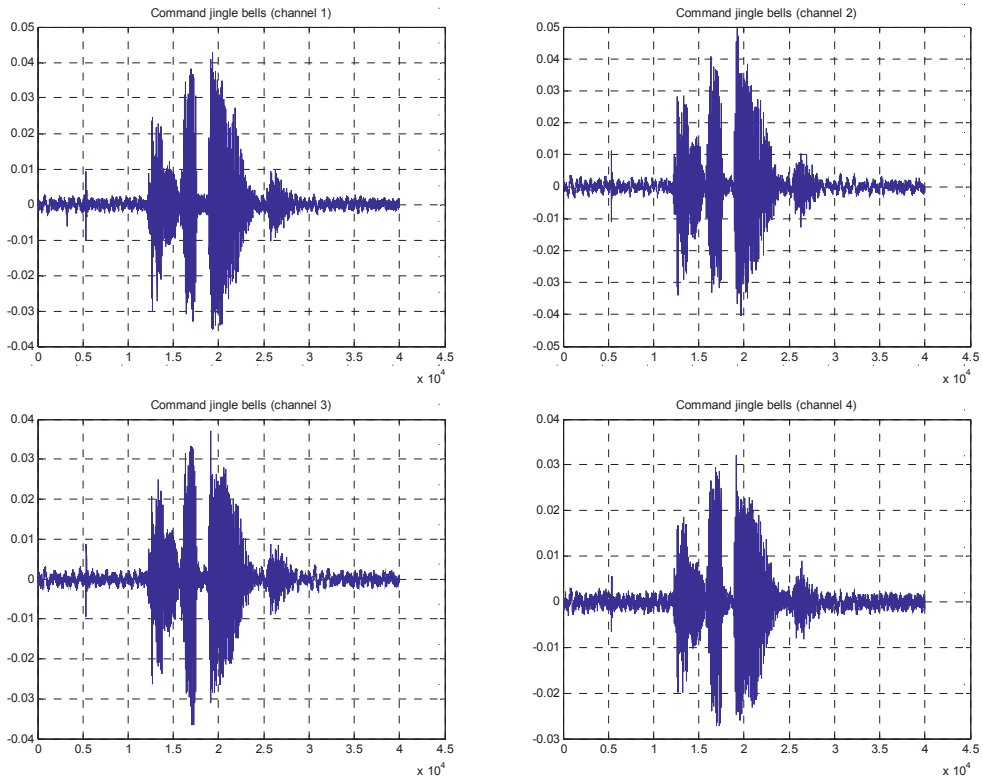

Figure 6 Command *jingle bells* in the four channels.

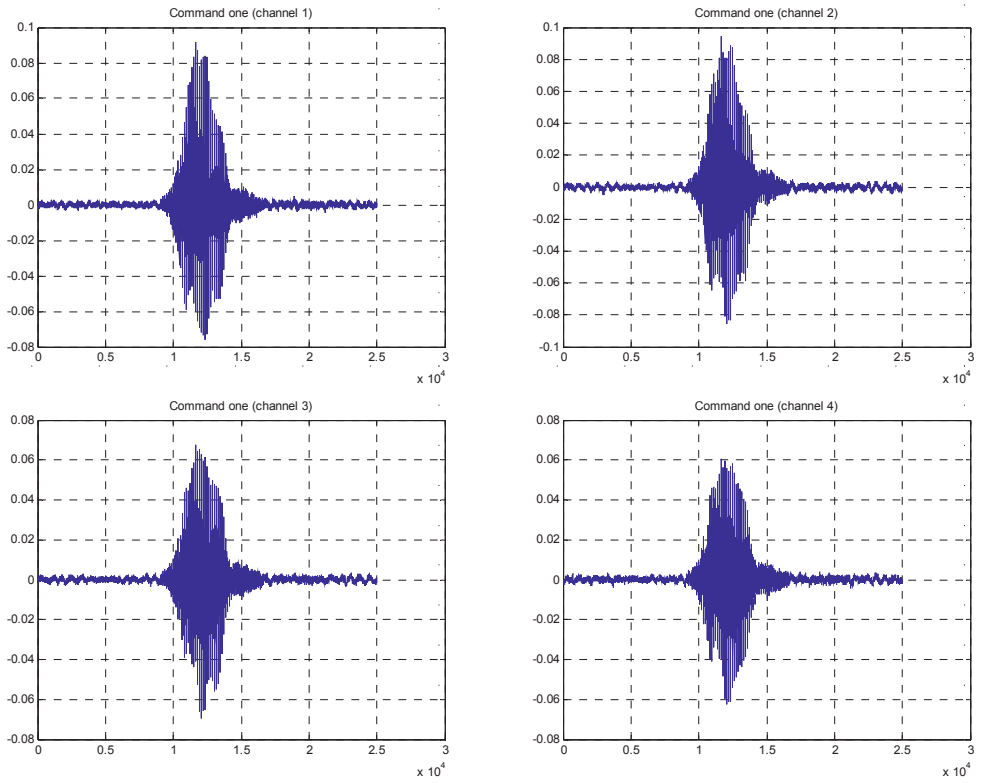

Figure 7 Command *one* in the four channels.