American Journal of Applied Sciences 6 (2): 290-295, 2009
ISSN 1546-9239
© 2009 Science Publications

Robust Speech Recognition Using Fusion Techniques and Adaptive Filtering

S.A.R. Al-Haddad, S.A. Samad, A. Hussain, K.A. Ishak and A.O.A. Noor
Department of Electrical, Electronic and Systems Engineering, Faculty of Engineering,
University Kebangsaan Malaysia, Bangi, Selangor, Malaysia

Abstract: The study proposes an algorithm for noise cancellation by using recursive least square (RLS) and pattern recognition by using fusion method of Dynamic Time Warping (DTW) and Hidden Markov Model (HMM). Speech signals are often corrupted with background noise and the changes in signal characteristics could be fast. These issues are especially important for robust speech recognition. Robustness is a key issue in speech recognition. The algorithm is tested on speech samples that are a part of a Malay corpus. It is shown that the fusion technique can be used to fuse the pattern recognition outputs of DTW and HMM. Furthermore refinement normalization was introduced by using weight mean vector to obtain better performance. Accuracy of 94% on pattern recognition was obtainable using fusion HMM and DTW compared to 80.5% using DTW and 90.7% using HMM separately. The accuracy of the proposed algorithm is increased further to 98% by utilization the RLS adaptive noise cancellation.

Key words: HMM, DTW, Zero crossing technique, RLS, word bounder

INTRODUCTION

Speech Recognition (SR) is a technique aimed at converting a speaker's spoken utterance into a text string or other applications. SR is still far from a solved problem. It is quoted that the best reported word-error rates on English broadcast news and conversational telephone speech are 10 and 20%, respectively[1]. Meanwhile, error rates on conversational meeting speech are about 50% higher and much more under noisy conditions[2].

Robustness is a key research in speech recognition for the past 50 years. The main issues in robustness are invariance to extraneous background noise and channel conditions as well as speaker and accent variations[3]. Recursive Least Squares (RLS) algorithm is used to improve the presence of speech in a background of noise. The RLS algorithm provides good performance for models with accurate initial information on a parameter or a state to be estimated[4]. In many applications of noise cancellation the changes in signal characteristics could be quite fast. This requires the utilization of adaptive algorithms, which converge rapidly. From this perspective the best choice is the RLS[5]. The beginning and end of a word should be detected by the system that processes the word after noise cancellation has been done.

Fusion pattern recognition is used such as with Dynamic Time Warping (DTW) and Hidden Markov Model (HMM). DTW is popularly used in speech recognition in the 70's and 80's[6,7] and HMM is popular after 90's until now[8]. Meanwhile, the fusion techniques started being used in the middle of 90's for complementing the benefits of each other[9,10]. There are a few types of fusion in speech recognition amongst them are HMM and Artificial Neural Network (ANN)[11] and HMM and Bayesian Network (BN)[12].

The algorithm is tested on Malay digit speech corpus. A hundred speakers were involved in this project each spoke with 10 repetitions for each digit. The Malay isolated digit are from 0-9 spoken as KOSONG, SATU, DUA, TIGA, EMPAT, LIMA, ENAM, TUJUH, LAPAN and SEMBILAN.

MATERIALS AND METHODS

The system begins with recording speech, RLS noise cancellation, end point detecting, framing, normalization, filtering, MFCC, weighting signal, time normalization, Vector Quantization (VQ) and labeling. Then HMM is used to calculate the reference patterns and DTW is used to normalize the training data with the reference patterns as in Fig. 1. In this paper Mel-

Corresponding Author: Syed Abdul Rahman A-Haddad, Department of Electrical, Electronic and Systems Engineering, Faculty of Engineering, University Kebangsaan Malaysia, 43600, UKM, Bangi, Selangor, Malaysia

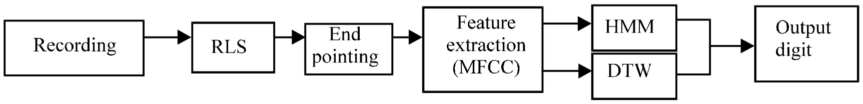

Fig. 1: Flowchart for fusion of Dynamic Time Warping (DTW) and Hidden Markov Model (HMM) with RLS noise canceller

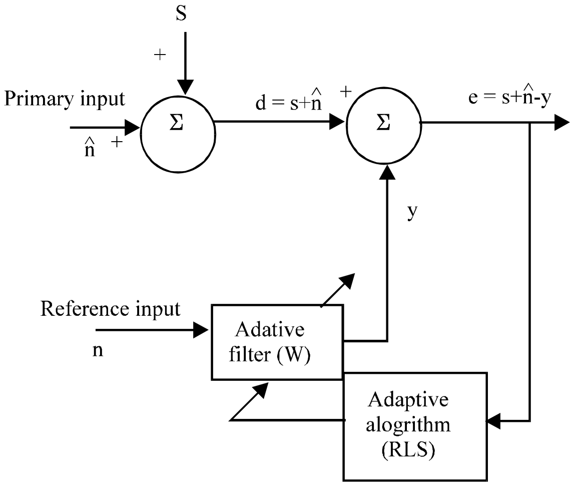

Fig. 2: Adaptive noise canceling

Frequency Cepstral Coefficient (MFCC) is chosen as the feature because of the sensitivity of the low order cepstral coefficients to overall spectral slope and the sensitivity properties of the high-order cepstral coefficients[13].

WAV file was recorded for 60 speakers. Each speaker says KOSONG, SATU, DUA, TIGA, EMPAT, LIMA, ENAM, TUJUH, LAPAN and SEMBILAN with a second pause for each number.

The RLS was used in preprocessing for noise cancellation as shown in Fig. 2[14]. The explanation for Fig. 2 is as follows:

n = Background noise of any type
$\hat{n}$ = Noise correlated to n
s = Speech signal
d = Desired signal
W = Optimum filter weight matrix
y = Output of adaptive process
e = Error signal in ideal case (clean speech)

Figure 3 shows the results of using the RLS adaptive filtering to the noisy signal. Figure 3a, shows the amplitude of the noisy speech and Fig. 3b shows the amplitude after processing using RLS.

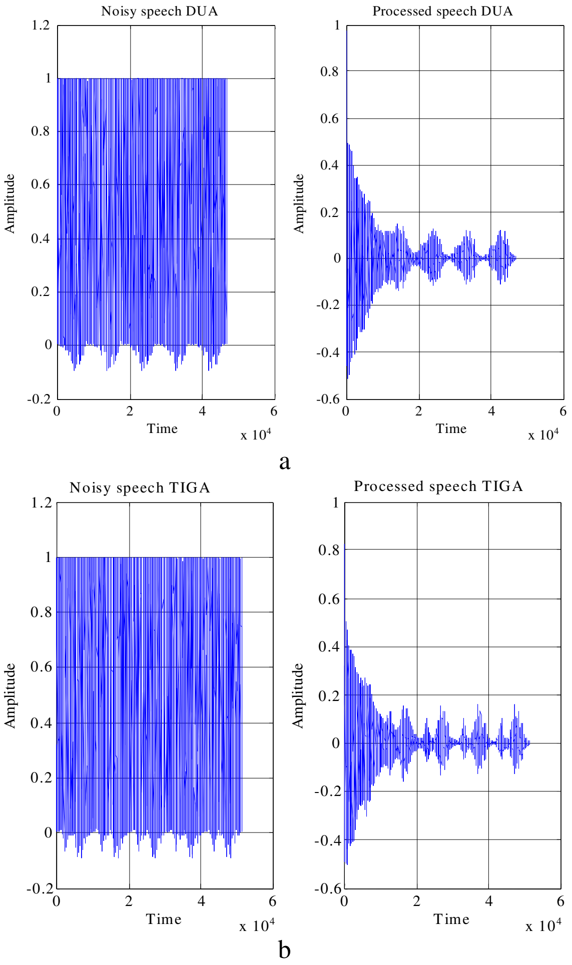

Fig. 3: (a): Noisy speech, (b): Signal processed by adaptive noise canceller

After getting the filtered noise speech sample, the first process is endpoint detection. For detection, two basic parameters are used: Zero Crossing Rate (ZCR) and short time energy. The energy parameter has been used in endpoint detection since the 1970's[15]. By combining with the ZCR, speech detection process can be made very accurate[16].

For labeling the segmented speech frame the zero crossing and energy were applied to the frame. Unfortunately it contained some level of background noise due to the fact that energy for breath and surround can quite easily be confused with the energy of a fricative sound[17].

As a result, this algorithm performs almost perfect segmentation for voice recoded by male speakers. For recoding done at noisy places, segmentation problem happens because in some cases the algorithm produces different values caused by background noise. This causes the cut off for silence to be raised as it may not be quite zero due to noise being interpreted as speech. On the other hand for clean speech both zero crossing rate and short term energy should be zero for silent regions.

Feature extraction: Mel Frequency Cepstral Coefficients (MFCC) is chosen because of the sensitivity of the low order cepstral coefficients to overall spectral slope and the sensitivity properties of the high-order cepstral coefficient[18]. Currently it is the most popular feature extraction method[18,19]. MFCC is produced after the recorded signal is pre-emphasized, framed and Hamming windowed. Then the signal is normalized and lowpass filtered. Lowpass filter is used to remove the potential artificial high frequencies appearing in their modulation spectrum due to transmission errors.

The Hamming window was calculated after getting the results from the endpoint process. The equation used is as follows:

$$w(n) = \frac{\alpha_w - (1-\alpha_w)\cos(2\pi n/(N_s - 1))}{\beta_w} \qquad (1)$$

where α_w is equal 0.54, meanwhile β_w, functions to normalized the energy through the operation so that the signal will not change. For the purpose of front end processing to obtain the desired frequency resolution on a Mel scale, the simple Fourier Transform (FT) is used. The average spectral magnitude for each amplitude coefficient is calculated as:

$$S_{avg}(f) = \frac{1}{N}\sum_{n=0}^{N} w_{FB^{(n)}|S(f)|} \qquad (2)$$

where the number of samples to get the average value is denoted as N, weighting function is denoted as $w_{FB^{(n)}}$ and magnitude of the frequency computed by the Fourier transform is denoted as $|_{s(f)|}$.

The cepstral coefficient is computed to minimize the non-information bearing variability from that amplitude via the following calculations:

$$c(n) = \frac{1}{N}\sum_{k=0}^{N} \log\left|S_{avg}(k)\right| \, e^{j\frac{2\pi}{N}kn}, 0 \le n \le N-1 \qquad (3)$$

where the average signal value in the kth is denote as S_{avg}.

Dynamic time warping (DTW): DTW is one of the main algorithms in this system for recognition after HMM. Due to the wide variations in speech between different instances of the same speaker, it is necessary to apply some type of non-linear time warping prior to the comparison of two speech instances. DTW is the preferred method for doing this, whereby the principles of dynamic programming can be applied to optimally align the speech signals. On the other hand, for detecting similar shapes with different phases, DTW has been used to calculate more robust distance for time series data. It can be used to measure similarity between sequences of different lengths. Because of these advantages many researchers use DTW such as for generic analysis and mining tasks on time series data, voice recognition and signature verification[20]. The distance metric used is a Euclidean distance for the cepstral coefficients over all frames after DTW is applied to align the frames optimally. The distance metric between frame i of the test word T_{MFCC} and frame j of the reference word R_{MFCC} is calculated as:

$$D_{ij} = \left(\frac{1}{p}\right)\sqrt{\sum_{k=1}^{p}\left(T_{MFCC}(i,k) - R_{MFCC}(j,k)\right)^2} \qquad (4)$$

This DTW algorithm has been tested with 80.5% correctness[21]. But for this fusion system the distance is calculated as:

$$D_{ij} = T_{MFCC}(i,k) - R_{MFCC}(j,k) \qquad (5)$$

for the purpose of processing one digit at a time. This distance will be used by decision fusion to process the weight mean vector for one digit.

Hidden markov model (HMM): HMM is typically an interconnected group of states that are assumed to emit a new feature vector for each frame according to an emission probability density function associated with that state. Viterbi algorithm is the most suitable for the estimation the parameters for HMM on the maximum likelihood criterion[22]. For HMM the expression is defined as λ = (A, B, π). A is denoted by a state transition probability matrix, B is denoted as output probability matrix and π denoted as initial state probability. The probability of the observation sequence p(o|λ) is given multidimensional observation sequences o, known as feature vectors.

For word-level HMM, the recognizer computes and compares all the p(o|λ_v) where (v = 1,2,…,W) and W is the digit word models. For left-to right, HMMs, p(o|λ_v) is computed using the Log-Viterbi algorithm as follows[23]:

for initialization,

$$\delta_1(j) = \log \pi_j + \log b_j(o_1) \text{ for } t = 1, 1 \le j \le N \quad (6)$$

for recursion,

$$\delta_t(j) = \max_{i=j-1,j}\left[\delta_{t-1}(i) + a_{ij}\right] + \log b_j(o_t) \quad \text{for } 2 \le t \le T,\ 1 \le j \le N \quad (7)$$

and for termination,

$$p(o|\lambda_v) = \max_{1 \le i \le N}\left[\delta_T(i)\right] \text{ for } t = T \quad (8)$$

The acronym used in the algorithm:

N = Number of states
T = Number of frames for feature vectors o = [o_1, o_2,…,o_T]
a_{ij} = State transition between i and j
A = {a_{ij}} are their N-by-N matrix
B = {logb_j (o_t)} is a N-by-T matrix in log output probability
δ_t (j) = Likelihood value at the time index t and state j

Fusion HMM and DTW: The pattern recognition fusion method used to fuse the results of DTW and HMM is weight mean vector. DTW measures the distance between recorded speech and a template, expanding or shrinking the temporal axis of the target to find the path or warping function which maximizes the similarity between the two speech signals. The distance of the signals is computed at each instant along the warping function. Meanwhile, HMM trains cluster and iteratively moves between clusters based on their likelihoods given by the various models. The weight mean vectors equation used is as follows:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} + \left|w_i - w_{i+1}\right| \quad (9)$$

which expands to,

$$\bar{x} = \frac{w_1 x_1 + w_2 x_2 + \ldots + w_n x_n}{w_1 + w_2 + \ldots + w_n} + \left|w_i - w_{i+1}\right| \quad (13)$$

Where:
w_1 = Query recognition rate in HMM test phase
w_2 = Query recognition rate in DTW test phase
x_n = Real time value of recorded speeches
$\bar{x}$ = Weight mean vector

For example if recognition percentage for HMM is h and for DTW is d for one digit, then in the fusion model after the query is recognized by DTW and HMM individually, the final percentage is calculated as follows:

$$\bar{x} = (((h * w_1) + (d * w_2)) / (w_2 + w_1)) + | h - d | \quad (14)$$

RESULTS AND DISCUSSION

We have evaluated the algorithm using the data described in the methodology section. The recognition algorithms HMM, DTW and DTW-HMM pattern recognition fusion is then tested for the percentage of accuracy. The test is limited to Malay digits from 0-9. Random utterance of digits is done and the accuracy of 100 samples is analyzed. The results obtained from the accuracy test is about 80.5% of accuracy for DTW and 90.7% for HMM and 94% for pattern recognition fusion. The results obtained are shown in Table 1.

Table 1: Comparison of digit recognition accuracy without using noise canceller.

Word	Accuracy HMM (%)	Accuracy DTW (%)	Accuracy Fusion (%)
Kosong	97	65	92
Satu	86	65	86
Dua	93	80	97
Tiga	86	65	86
Empat	86	100	100
Lima	87	75	92
Enam	86	100	100
Tujuh	86	65	86
Lapan	100	95	100
Sembilan	100	95	100
Average	90.7	80.5	94

Table 2: Comparison digit recognition accuracy with noise canceller

Word	Accuracy HMM (%)	Accuracy DTW (%)	Accuracy fusion (%)
Kosong	98	82	99
Satu	90	82	94
Dua	94	98	98
Tiga	92	82	97
Empat	92	100	98
Lima	93	94	99
Enam	92	100	100
Tujuh	91	84	98
Lapan	100	96	99
Sembilan	100	96	99
Average	94.2	91.4	98.1

Meanwhile for robustness, the speech is first filtered by using RLS noise cancellation, the results obtained are as shown in Table 2. Noise cancellation increases the accuracy for HMM, DTW and Fusion to 94.2, 91.4 and 98.1%, respectively.

CONCLUSION

This research has shown a speech recognition algorithm using MFCC vectors to provide an estimate of the vocal tract filter. DTW and HMM are the two recognition algorithms used. DTW is used to detect the nearest recorded voice. Meanwhile HMM is used to emit a new feature vector for each frame according to an emission probability density function associated with that state. The results showed a promising speech recognition module as tested on a Malay digit database. This paper has shown that the fusion technique can be used to fuse the pattern recognition outputs of DTW and HMM. Furthermore it also introduced refinement normalization by using weight mean vector to get better performance with an accuracy of 94% for pattern recognition fusion HMM and DTW. This can be compared to the accuracy for DTW and HMM, which is 80.5 and 90.7%, respectively. The accuracy is further increased after RLS noise cancellation to 98.1% for the fusion technique.

ACKNOWLEDGMENTS

This research is supported by the following research grants: Fundamental Research Grant Scheme, Malaysian Ministry of Higher Education, FRGS UKM-KK-02-FRGS-0036-2006 and E-Science 01-01-02-SF0374, Malaysian Ministry of Science, Technology and Innovation.

REFERENCES

1. Le, A., 2003. Rich Transcription 2003: Spring speech-to-text transcription evaluation results, Proc. RT03 Workshop, 2003. May 19-20, 2003, Boston, MA, USA. <http://www.nist.gov/speech/tests/rt/rt2003/spring/presentations/rt03s-stt-results-v9.pdf>. Accessed date: Oct 25, 2007.
2. Le, A., J. Fiscus, J. Garofolo, M. Przybocki, A. Martin, G. Sanders and D. Pallet, 2007. The 2002 NIST RT evaluation speech-to-text results. In Proceeding RT02 Workshop, May 7-8, 2002, Vienna, Va, USA. http://www.nist.gov/speech/tests/rt/rt2002/presentations/rt02_stt_results_v5.pdf. Accessed date: Oct 25, 2007.
3. Huang, X.D., A. Acero and H.W. Hon, 2001. Spoken Language Processing. 1st Ed., Englewood Cliffs, Prentice-Hall NJ.
4. Park, J.H., Z.H. Quan, S. Han and W.H. Kwon, 2008. New recursive least squares algorithms without using the initial information. IEICE Trans. Commun., E91-B: 968-971.
5. Vijayakumar, V.R. and P.T. Vanathi, 2007. Modified adaptive filtering algorithm for noise cancellation in speech signals, electronics and electrical engineering. Kaunas Technol., 74: 17-20.
6. Sakoe, H. and S. Chiba, 1975. Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. ASSP., 26: 43- 49.
7. Zhu, Y. and D. Shasha, 2003. Warping indexes with envelope transforms for query by humming. Proceeding of the ACM SIGMOD International Conference on Management of Data (SIGMOD 2003). June 9-12, 2003, San Diego, California, pp: 181-192.
8. Rabiner, L.R., 1989. A tutorial on hidden markov models and selected applications in speech recognition, Institute of Electrical and Electronics Engineers Inc. (IEEE) Transactions Speech Audio Process., 2: 257-285.
9. Rabiner, L.R. and M.R. Sambur, 1975. An algorithm for determining the endpoints of isolated utterances. Bell. System. Tech. J., 54: 297-315.
10. Rabiner, L.R. and R.W. Schafer, 1978. Digital Processing of Speech Signals, Prentice-Hall Inc., NY, USA.
11. Rabiner, L.R., 1989. A tutorial on hidden markov models and selected applications in speech recognition, Institute of Electrical and Electronics Engineers Inc. (IEEE) Trans. Speech Audio Process., 2: 257-285.
12. Thian, N.P.H., S. Bengio and J. Korczak, 2002. A Multi-Sample Multi Source Model For Biometric Authentication, Institut Dalle Molle d'Intelligence Artificielle Perceptive (IDIAP) Research Report, pp: 02-14.
13. Kim, T.Y. and H. Ko, 2005. Bayesian fusion of confidence measures for speech recognition, Institute of Electrical and Electronics Engineers Inc. (IEEE) Signal Process. Lett., 12 (12).
14. Ifeachor, E.C. and B.W. Jervis, 2004. Digital Signal Processing-A practical Approach, Pearson Education, Delhi, India.
15. Analog Devices Inc., 1992. Digital Signal Processing Applications using the ADSP-2100 Family. Vol. 2. Prentice Hall.

16. Gold, B. and N. Morgan, 2000. Speech and Audio Signal Processing. 1st Edn. John Wiley and Sons, New York, USA, pp: 537.
17. Al-Haddad, S.A.R., S.A. Samad and A. Hussain, 2006. Automatic digit boundary segmentation recognition. MMU International Symposium on Information and Communications Technologies (M2USIC) 2006, 16-17 November 2006. Petaling Jaya, Selangor, pp: 212-217.
18. ESTI, 2002. Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-end Feature Extraction Algorithm, Compression Algorithm. European Telecommunications Standards Institute (ETSI) Standard Document, ES 201 108.
19. Zhu, Q. and A. Alwan, 2000. On the use of variable frame rate analysis in speech recognition. Proc. IEEE ICASSP, Turkey, 3: 1783-1786.
20. Chu, S., E. Keogh, D. Hart and M. Pazzani, 2002. Iterative deepening dynamic time warping for time series. In: Proceeding of SIAM International Conference on Data Mining, Hyatt Regency Crystal City, Ronald Reagan Nattional Airport Arlington, VA, US. April 11-13, 2002. Society for Industrial and Applied Mathematics (SIAM), pp: 195-212.
21. Al-Haddad, S.A.R., S.A. Samad and A. Hussain, 2007. Automatic recognition for malay isolated digits. The 3rd International Colloquium on Signal Processing and its Applications (CSPA 2007), March 9-11, 2007. Melaka, Malaysia, pp: 97-101.
22. Yoshizawa, S., N. Wada, N. Hayasaka and Y. Miyanaga, 2002. Scalable architecture for word HMM-based speech recognition and VLSI implementation in complete system. IEEE Trans. Circuits Syst., 1: 70-78.
23. Young, S., G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason and D. Povey, V. Valtchev and P. Woodland, 2002. The HTK Book. Version 3.2. Cambridge University, UK, CUED Publications, UK.