ISSN 0973-3302

2010

PROCEEDING OF NATIONAL SYMPOSIUM ON ACCOUSTIC

Speech, Hearing and Music

VOLUME-37 NUMBER 1 2010

www.nsa2010.gpgcrishikesh.com

Journal of Acoustical Society of India

The Refereed Journal of the Acoustical Society of India (JASI)

The journal of Acoustical Society of India is a refereed journal of the Acoustical Society of India (ASI). The ASI is a non-profit national society founded in 31st july,1971. The primary objective of the society is the advance the science the science of acoustics by creating an organization that is responsive to the needs of scientists and engineers concerned with acoustics problems all around the world.

Manuscripts of articles, technical notes and letter to the editor should be submitted to the Chief Editor. Copies of articles on specific listed above should also be submitted to the respective. Associate Scientific Editor. Manuscripts are refereed by at least two referees and are reviewed by Publication Committee (all editors) before acceptance. On are refereed by at least two referees and are reviewed by Publicatin Committee (all editors) before acceptance. On acceptance revised articles with the text and figures scanned as separate files on a diskette should be submitted to the Editor by express mail Manuscripts of articles must be prepared in strict accordance with the author instructions.
All information concerning subscription new books journals conferences etc. Should be submitted to Chief Editor

Acoustics Section, National Physical Laboratory, Dr. KS Krishnan Road New Delhi 110 012
Tel: +91.11.4560.9319. Fax:+91.11.4560.9310, e-mail:mahavir@nplindia.org

Copyright @2007, Acoustical Society of India
ISSN 0973-330
Printed at Alpha Printers BG-2/38C Paschim Vihar, New Delhi-110063 Tel:9811848335, JASI is sent to ASI members free of charge.

Sound Transmission through Building Enclosures

The article deals with sound transmission through building enclosures including different forms of predominantly lightweight wall construction. It gives guidance on how walls comprising a number of separate element can be assessed. Sound transmission from outside to inside the reader should also see BS EN 12345-4:2000 if they are concerned with containment of sound within a building.

The performance of a wall or roof has to be considered in terms of the internal spaces. The aim is to provide a building envelope that gives the required sound pressure levels within a room or other internal space. The noise level within a room will depend on the amount of sound energy transmitted through the wall and interreflection of sound inside the room. The room effect is usually determined by the amount of sound absorbing material in the room. Sound transmission of an assembly of components can be calculated provided the wall can be analysed as discrete areas for each of which the Sound Reduction Index is known.

This applies to windows in walls and collections of windows but note that sound transmission through interface components such a joining mullions between windows may not be known.

Sound transmission through a whole wall Is established by calculating an apparent sound reduction index (SRI) for the wall. This is used to determined the difference in sound between the outside and inside. The procedure is to calculate the sound power reduction for each element of the wall. The total sound power reduction can then be calculated and converted to an apparent reduction index.

When sound of intensity 1W/m^2 falls on a wall the sound power (in watts) transmitted by an element is given by:

$W_i = S_i \; 10$

Where

S_i is the area of an element (m^2)

R_i is the sound reduction Index of that element (dB)

(Mahavir Singh)

SAVING OF COMPUTATIONAL TIME AND REJECTION OF NON-VOCABULARY WORD IN DTW AND LPC BASED SPOKEN WORD RECOGNIZER

Mohammad Saleem[1], Shaghaf M. Ansari[2] and Safdar Tanweer[3]

[1]IT Analyst, Tata Consultancy Services Ltd, Pune-411 057, India

Email: mdsaleem@gmail.com

[2]M.Tech. Student, Department of Computer Science, Jamia Hamdard, New Delhi-110062

[3]Department of Computer Science, Jamia Hamdard, New Delhi-110062, India

ABSTRACT: Real-time response, accuracy, and rejection of non-vocabulary word etc. are the major problems in automatic speech recognition system. In the present study, an attempt has been made to solve the problem of real-time response of dynamic time warping (DTW) based automatic speech recognizer by saving the computational time and rejecting the non-vocabulary word. To speed up the DTW computation of the speech recognizer, a method for aborting the DTW computation in the case of non-matching word is described, which is based on the automatically calculated word rejection threshold levels corresponding to each and every LP Coefficients. These threshold levels are automatically calculated on the basis of the initial recognition performance of the chosen vocabulary of words as a part of the training process. The threshold corresponding to individual LP Coefficients ensures that a sufficient number of unlike patterns are rejected which results in saving the computational time. During the recognition process, it has been observed that most of the time the DTW computation of one or two LP Coefficients are sufficient to abort the DTW computation in the case of dissimilar word. The described method offers more than 84% of actual savings of time in computing the recurrence relation as compared to the conventional DTW methods without degrading the recognition accuracy [1]-[8].

KEYWORDS: Isolated word, Speech recognition, Dynamic time warping, LPC

1. Introduction

It has long been observed that the human pronunciation of a word at different occasions may differ in some of their acoustic characteristics; the most commonly observed changes are in speaking intensity and word duration. However, these variations affect the recognition performance of automatic speech recognition system in terms of the poor recognition score.

In order to solve the problem of poor recognition score of the isolated word speech recognition system due to variation in speaking rate, the linear and non-linear methods of time alignment are most commonly used. It has long being realized that the non-linear method of time normalization gives better recognition score than the linear one. The DTW is a nonlinear popular algorithm for measuring the similarity between two sequences which may vary in time or speed. This algorithm is being used in recognizing the spoken word for the last several decades as it is one of the prominent techniques to accomplish the task of non-linear time alignment in speech recognition systems [1]-[8]. In this technique, a test template is stretched or compressed according to the reference template. More clearly, DTW process nonlinearly expands or contracts the time axis to match the same phoneme positions between the input speech and reference templates. In spite of the fact that the computation of recurrence relation for realizing the non-linear time alignment (DTW) faces a serious problem of high computational time, the DTW is widely used in the small-scale embedded speech recognition systems such as those embedded in many voice controlled appliances. The reason for this is owing to the simplicity of the hardware implementation of the DTW algorithm, which makes it suitable for many mobile devices. Additionally, the training procedure in DTW is very simple and fast, as compared with the HMM and ANN rivals [9]. However, the DTW technique is reliable and computationally expensive and can be easily implemented in any hardware or software based recognizer.

DTW being a computationally expensive and practically useful technique for the development of fully automatic and standalone speech recognition system, an attempt has been made to speed up the computational time by rejecting the dissimilar and non-vocabulary words of the DTW recognizer.

2. Feature Extraction

Among the several feature being extracted from the voice samples for speech recognition, the linear prediction coefficient (LPC) is the most common and useful feature [10]-[12]. It has been widely accepted that linear prediction coefficient is an analytically tractable model and provides a good approximation to the vocal tract spectral envelop [13]. The linear prediction analysis procedure is applied to each short interval of time, known as frame. Within a frame, the weights used to compute the linear combination are found by minimizing the mean-squared prediction error. However, the extracted LPCs from each frame result in a time varying filter representing the activity of human speech production organ [14]-[15]. Linear Prediction can also be viewed as a redundancy removal procedure where information repeated in an event is eliminated; therefore, there is no need of a data if it can be predicted.

In the present study, the 12-LP coefficients were computed from the .wav files of spoken words stored on a multimedia PC. In this way a database consisting of 10 utterances of English and Hindi digits (zero through nine), control words, and other words quite frequently used in daily life etc. spoken by ten male and female speakers were created for the vocabulary of 100 words and used for performing the recognition experiments.

3. Dynamic Time Warping Algorithm

In the present system, dynamic programming approach has been used in order to minimize the variation in speaking rate effectively [8]. It is well known that the speech can be expressed by appropriate feature extraction as a sequence of feature vectors. Suppose we have two series of time sampled speech patterns A and B of length I and J respectively for computing the distance between the patterns A and B, we apply the dynamic time warping algorithm as follows.

Let $A = a_1, a_2, \ldots, a_i, \ldots, a_I$ and $B = b_1, b_2, \ldots, b_j, \ldots, b_J$, where a_i and b_j are time- sampled feature vectors of A and B respectively, and $d(i, j) = a_i - b_j$ be distance between a_i and b_j. By calculating the recurrence relation,

$$g(i,j) = d(i,j) + \mathrm{Min} \begin{array}{l} g(i, j-1) \\ g(i-1, j) \\ g(i-1, j-1) \end{array} \qquad (1)$$

up to (I, J) with initial condition,

$$g(1, 1) = d(1, 1) \qquad (2)$$

The similarity between A and B is obtained as,

$$S(A, B) = g(I, J) / (I + J) \qquad (3)$$

Actually, the combination of a_i and b_j in Eq.(1) is restricted within the domain called adjustment window defined by

$$W = \{|(I, J)|, |I - J| \leq r\} \qquad (4)$$

Where, a positive integer r is chosen so that the timing variation is practically limited within Eq. (4). Introduction of the restriction (4) reduces the amount of computation considerably [8]. The domain in which the dynamic programming equation must be calculated is also specified by $1 \leq i \leq I$, $1 \leq j \leq J$, and $j-r \leq i \leq j+r$ (adjustment window) [3]. The programming of the present DTW algorithm for measuring the similarity between test and reference template is conveniently simplified with an assumption that the short template is test template and long template is reference template temporarily and window size $|I-J|$ was taken as $|I-J|+1$. More clearly, the computation sequence of recurrence relation g(i, j) within the boundary conditions are as follows.

$g(1,1), g(1,2), \ldots, g(1, I-J+1)$

$g(2,1), g(2,2), \ldots, g(2, I-J+1)$

$g(3,2), g(3,3), \ldots, g(3, I-J+1)$

$g(i,j-1), g(i,j), \ldots, g(i, I-J+1)$

----

up to g(I,J) without considering the existence of accumulative distances g(i,0) and g(0,j).

4. Implementation Issues

The above mentioned DTW algorithm was implemented for the recognition of spoken word based on the techniques of saving the computational time and rejecting the non- vocabulary word [16]. The idea of the DTW technique is to match a test input represented by a multi-dimensional feature vector $A = a_1, a_2, \ldots, a_i, \ldots, a_I$ with the reference template $B = b_1, b_2, \ldots, b_j, \ldots, b_{J,}$ [3]-[5]. The aim of dynamic time warping is to find the similarity S(A,B) such that it gives the least cumulative difference between the compared templates. To ease the computation of each comparison of the test and reference templates, the smaller template is always taken as test input and longer template as reference template temporarily. This provides the idea of computation of cumulative distance g(i,j) always in forward direction.

To speedup the distance computation by terminating the computation of accumulative distance g(I, J) of unlikely patterns, the cumulative distance rejection thresholds have been used. The accumulated distances g(I, J) using recurrence relation is sequentially computed starting from 1st LP coefficient to 12th LP coefficient followed by their successive additions (cumulative distance). If we denote the cumulated distance as D_l, and the rejection threshold as T_l corresponding to lth LP coefficient, then if

$D_l > T_l \quad (l=1, 2, \ldots, L)$

The computation is aborted at lth LP coefficient and the test word corresponding to the reference pattern is rejected as a candidate for rejection. The process for the computation of T_l is as follows.

During the training process, two utterances of the whole vocabulary of words are recorded out of which one utterance is treated as the permanent reference template and the whole vocabulary of word of the other utterance is tested for their recognition in a sequential order as spoken. If all the words are accurately recognized then their corresponding minimum accumulative distances g(I, J) for all the 12 LP coefficients are stored as a row of a matrix. More clearly, the number of accumulative distance of LP coefficients represents the columns of the matrix and the number of rows depends on the size of the vocabulary of words as follows.

$$\begin{array}{cccccccc}
g_{11}(I, J) & g_{12}(I, J) & - & - & g_{1l}(I, J) & - & - & g_{1L}(I, J) \\
g_{21}(I, J) & g_{22}(I, J) & - & - & g_{2l}(I, J) & - & - & g_{2L}(I, J) \\
- & - & - & - & - & - & - & - \\
- & - & - & - & - & - & - & - \\
g_{n1}(I, J) & g_{n2}(I, J) & - & - & g_{nl}(I, J) & - & - & g_{nL}(I, J) \\
- & - & - & - & - & - & - & - \\
- & - & - & - & - & - & - & - \\
g_{N1}(I, J) & g_{N2}(I, J) & - & - & g_{Nl}(I, J) & - & - & g_{NL}(I, J)
\end{array}$$

Where, I varies from 1 to 12 and represents the minimum accumulative distances g(I, J) corresponding to 12 LP coefficients, and n varies from 1 to a maximum value of 100 depending upon the size of the vocabulary of word.

In order to tolerate the intra-speaker variations or variations due to microphone position etc., the maxima of each column of the above stored matrix is finally stored in a single row as follows.

$$g_{M1}(I, J)\ g_{M2}(I, J)\ -\quad -\quad g_{Ml}(I, J)\ -\quad -\quad g_{ML}(I, J)$$

These maxima values are being used for computing the rejection threshold as follows.

$$T_1 = g_{M1}(I, J)$$
$$T_2 = T_1 + g_{M2}(I, J)$$
$$T_3 = T_2 + g_{M3}(I, J)$$
$$\dots\dots\dots\dots\dots$$
$$\dots\dots\dots\dots\dots$$
$$T_{12} = T_{11} + g_{M12}(I, J)$$

Since these rejection thresholds are speaker as well as vocabulary dependent that is why it is highly reliable to reject the dissimilar words. It is not unusual for the most majority of incorrect words to be eliminated by the rejection thresholds. However, this rejection threshold saves more than 84% of actual computational time by allowing more templates to go to termination in DTW computation without degrading the rejection accuracy. Hence, two more additional features, namely; rejection of non-vocabulary (foreign) word, and reduction of computation time were successfully incorporated in the present recognizer.

The last major step in the pattern-recognition model of speech recognition is the decision rule which chooses which (reference) pattern (or patterns) most closely matches with the unknown test pattern. It is based on the cumulative distances obtained from the recurrence relation of the test and reference patterns as follows.

$$\text{Distance} = \sum_{l=1}^{12} S(A_l,B_l)$$

Where, $S(A_l,B_l)$ is the similarity between A_l and B_l (test and reference) time-sampled speech patterns with respect to the l^{th} LP coefficient.

5. Experimental Results

The recognition performance of the speech recognizer was evaluated on the basis of different aspects of test conditions such as size of the vocabulary of words, spoken languages, male and female speakers, and socioeconomic background of the speakers etc. The overall variation of recognition score was found to be 94% to 100% for a chosen vocabulary of 100 words.

The described method offers more than 84% of actual savings of time in computing the recurrence relation as compared to the conventional DTW methods without degrading the recognition accuracy [1]-[8]. However, an additional facility for rejecting the non-vocabulary word is also included in the recognizer.

Acknowledgements

The authors are grateful to Prof. Abdul Mobin, National Physical Laboratory, New Delhi and Prof. M. Afshar Alam, Jamia Hamdard University, New Delhi for their expert guidance and encouragement.

References

[1] Rabiner L, Rosenberg A and Levinson S, “Consideration in dynamic time warping algorithms for discrete word recognition”, IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. ASSP-26, pp.575-582, 1978.

[2] Myer C, Rabiner L and Rosenberg A, “Performance tradeoffs in dynamic time warping algorithms for isolated word recognition”, IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. ASSP-28, pp.623-635, 1980.

[3] Sakoe H & Chiba S, “Dynamic programming algorithm optimization for spoken word recognition”, IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. ASSP-26, pp.43-49, 1978.

[4] Tappert C and Das S, “Memory and time improvement in a dynamic programming algorithm for matching speech patterns”, IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP-26, pp.583-586, 1978.

[5] Sakoe H and Chiba S, “Recognition of continuously spoken words based on time-normalization by dynamic programming”, J. Acous. Soc. Japan, Vol.27, No.9, pp.483-500, 1971.

[6] Silverman H F, and Morgan D P, “The application of dynamic programming to connected speech recognition”, IEEE ASSP Magazine, pp.7-25, 1990.

[7] George M. White and Richard B. Neely, “Speech recognition experiments with Linear prediction, bandpass filtering, and dynamic programming”, IEEE Trans. on ASSP, Vol. 24, No.2, pp.183-188, 1976.

[8] Sakoe H and Chiba S, “A dynamic programming approach to continuous speech recognition”, Proc. Int. Congress on Acoustics, Budapest, Paper 20, p.C13, 1971.

[9] Trentin E, “Robust combination of neural networks and hidden Markov models for speech recognition”, Ph.D Thesis, University of Florence (Italy), 2001.

[10] Rabiner L, Juang B, “Fundamentals of Speech Recognition”, Prentice Hall Signal processing Series, 1993.

[11] Furui S, “Digital Speech Processing, Synthesis and Recognition”, Marcel Dekker, Inc., New York, 2000.

[12] J.D Markel and A.H Gray, “Linear Prediction of Speech”, Berlin, Germany: Springer-Verlag, 1976.

[13] Wai, C Chu, “Speech Coding Algorithms: Foundation and Evolution of Standardized Coders”, John Willey & Sons, 2003.

[14] Einarsson T and Thorsteinsson J, “Speech Linear Prediction and Synthesis”, University of Maryland, Project 2 for ENEE624 – Fall, 2002.

[15] Islam T, “Interpolation of Linear Prediction Coefficients for Speech Coding”, ME Thesis, McGill University, Montreal, Canada, 2000.

[16] Mobin Abdul and Agrawal S.S, “Role of filter-bank features in DTW based word recognition for saving the computational time and rejecting the non-vocabulary word”, Proceedings of International Conference on Information Systems, Analysis and Synthesis (ISAS-2001) and World Multi-conference on Systemics, Cybernetics and Informatics (SCI-2001), Orlando, Florida (USA), Vol. XIII, July 22-25, 2001.