Effect of Regional Variants on Isolated Word Recognition in India

Mohd Saleem
Research Scholar, Dept of Computer Science, Jamia Millia Islamia, New Delhi-25
mdsaleem@gmail.com

K Mustafa
Associate Professor, Dept of Info Technology AHB Talal University, Ma'an JORDAN
kmfarooki@yahoo.com

RA Khan
Lecturer, Dept of Computer Science, Jamia Millia Islamia, New Delhi
khanraees@lycos.com

Pratibha Agrawal
Research Scholar, Dept of Computer Science, Jamia Millia Islamia New Delhi-25

Abstract

The variability of speech between the speakers belonging to different region is an obstruction for the robustness in the speech recognition. Several techniques are used to normalize different speakers; one of them is to deal with the problem of regional variation. In this paper we discuss the problem of whether the regional variation affects the accuracy of speech recognizer and whether we can improve the performance of the recognizer through knowledge of the regional origin of the unknown speaker.

The fundamental concept in developing such robust system may be to train the recognizer by clustering test speakers into dissimilar dialectal regions. We have studied and analyzed through number of experiments to authenticate the enhancement in the overall accuracy of the recognizer. A progressive enhancement in recognition accuracy may be achieved for a simple pattern matching recognizer while it may not significantly improve the recognition accuracy for other stochastic modeling techniques.

Keywords: Regional variants, pattern matching, stochastic modeling, dialectal region.

1. Introduction

The modern day's speech recognizers are still awaiting 100% recognition accuracy due to the differences in positioning, tension and movement of laryngeal anatomical parameters which lead to the different accent for the speakers belonging to different region. The regional geographical area or a social grouping makes a great difference on the accent in pronunciation by a community of people. The differences in phonetic transcriptions and the acoustic correlates of speech, including formants and their trajectories, pitch trajectory, pitch nucleus and duration parameters are the important factors which affects the accent [1]. One way of improving the recognition accuracy is to exploit language specific phenomena like dialectal influences [2]. Here we investigated the influences of regional variants in English spoken words since the real dialects are easy to spot and can be treated as own languages.

2. Conceptual Background

Concerning the situations in which the spoken word recognizer is to be used, it is evident that speakers do not speak like a "pronunciation dictionaries" but rather use a controlled version of their everyday life speech [3][4][5]. The number of questions arises here to achieve robustness; to find out the regional variants which are making impact on isolated word recognition? Is it possible to improve the recognition accuracy by considering the weak regional variants during the recognition process?

Since the stochastic modeling based recognizer are already so much robust that they does not provide any enhancement by using regional variants rather the accuracy degraded due to the extension of search area in the recognition process. However, it has been observed that the DTW is widely used in the small-scale embedded-speech recognition systems such as those embedded in many voice controlled appliances [6]. 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 [7]. However, the DTW technique is reliable and computationally expensive and can be easily implemented in any hardware or software based recognizer.

The problem in speech recognition occurs due to the number of reasons depending on the regional variations like speaking rate, word duration, accent and pitch etc. Some of the English digit spoken words, changed due to the regional effect, have shown below:

Spoken Word	Error Word	Phonetic Error
War	Waar	/a/ to /ae/
Victory	Bictory	/v/ to /b/
Entirely	Yentirely	/e/ to /y/
Stop	Stoap	/o/ to /oa/
Morning	Marning	/o/ to /a/
High	I	Consonant Removal
Zoo	Soo	/z/ to /s/

It has been also observed that the speech recognition is always depending on the vowels (/a/, /e/, /i/, /o/, /u/) and the role of the consonant is very least on the recognition performance [8][9].

3. Feature Extraction

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. 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. Linear Prediction can also be viewed as a redundancy removal procedure where information repeated in an event is eliminated; therefore, there is no need for a data if it can be predicted [10].

However, the 10-LP coefficients were computed from the wave files of spoken English digits by several speakers stored on a multimedia PC [11]. In this way a database consisting of 10 utterances of English spoken words by several speakers belonging to four broad region of the country were created and used for the recognition experiments. The South India, North India, Bihar and West Bengal were considered as four regions.

Spoken Word Recognizer

In this study, the dynamic time warping (DTW) based recognizer is used to perform the speech recognition experiments. The dynamic programming approach has been used in order to minimize the speaking rate variation effectively as follows:

Suppose we have two series of time sampled speech patterns A and B, of length I and J respectively. 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.

To align the two time-sampled series using DTW, we construct an I-by-J matrix where the (i^{th},j^{th}) element of the matrix contains the distance $d(a_i,b_j) = a_i-b_j$ between two points a_i and b_j. The alignment of two time series samples can be found very efficient using dynamic programming to evaluate the following recurrence which defines the cumulative distance g(i,j) as the distance d(i,j) found in the current cell and the minimum of the cumulative distances of the adjacent elements:

$$g(i,j) = d(a_i,b_j) + \min \begin{cases} g(i-1,j-1) \\ g(i-1,j) \\ g(i,j-1) \end{cases} \quad \ldots (1)$$

The above recurrence relation is calculated up to (I,J) with initial condition

$$g(1,1) = d(a_1,b_1) \quad \ldots (2)$$

The similarity between A and B is obtained as

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

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

$$W = \{ (i,j) \mid |i-j| \leq r \} \quad \ldots (4)$$

Where a positive integer r is chosen so that the timing variation is practically limited within (4). Introduction of the restriction (4) reduces the amount of computation considerably.

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 the unknown test pattern. It is based on the cumulative distances obtained from the recurrence relation of the reference and test patterns as follows:

$$\text{Distance} = \sum_{L=1}^{10} S(A_L,B_L)$$

Where $S(A_L,B_L)$ is the similarity between AL and BL (reference and test) time-sampled speech patterns with respect to the output of the L^{th} LP coefficient and 10 is the order of linear prediction coding. The

reference template having minimal distance with the test template is the recognized template [8].

5. Experiments

The numbers of test experiments were conducted for the speakers belonging to different states of India. Since the differences in positioning, tension and movement of laryngeal anatomical parameters lead to the different accent for the speakers belonging to different region; a set of dictionaries containing regionally clustered variants was built by using the dialect-specific transcribed pronunciation forms and clustered into 4 broad dialectal region of India.

The effect of regional variation on the accuracy of speech recognizer was improved through the knowledge of the regional origin of the unknown speaker. One more way of improving the recognition accuracy was explored by adding the regional information in the training and recognition phase to mask the effect of language specific phenomena like dialectal influences. The fundamental concept in developing such robust system was to train the recognizer by clustering test speakers from dissimilar dialectal regions.

5. Conclusion

The regional variants like speaking style, word duration, accent and tone of the speakers make a massive impact on the recognition performance. Here we have investigated the effect of the regional variants on isolated word recognition of English spoken words. In this research work, we tried to explore how the regional variants can be effectively masked for better recognition of spoken words. We reached to the conclusion that for the stochastic modeling based recognizer the regional variants are no more issue because of an already existing good robustness of the recognizer. While for simple pattern matching based recognizer the consideration of the regional variants will be beneficial for the improvement of recognition accuracy.

6. Future Work

The future work will be considered to use different modeling for different regions to improve the recognition accuracy for isolated word recognizer based on wavelet transform. By creating different database dictionary for different regions of India, we can train the system for speakers belonging to different regions to get the improved accuracy. Also it may gives another drawback of more processing due to huge database created by making the same corpus of speech for all four or more than four basic regions. To avoid such circumstance, the separate modeling for all the acoustically differentiable regions may be much better technique. In this technique, during the recognition phase the utterance is categorized to the model from where the speaker acoustically belongs to and then the appropriate model may be used for the recognition purpose.

7. Acknowledgements

The authors are grateful to Dr. Abdul Mobin, Scientist-in-Charge CEERI, Delhi Centre for his constant encouragement, kind support and valuable guidance.

References

[1] Yan Q, Vaseghi S, “Analysis, Modelling and Synthesis of Formants of British, American and Australian Accents”, ICASSP, Vol I, pp. 712-715, 2003.

[2] Burger S, Draxler C, “Identifying dialects of German digit strings”, Proc. Int. Conf. on languages source and evaluation, Spain, 1998.

[3] Gerosa M, Giuliani D, “Preliminary investigation in automatic recognition of English sentences uttered by Italian children”, proc InSTIL/ICALL, Venice, 17-19 June 2004.

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

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

[6] Mobin Abdul, Saleem M, “Saving of computational time and rejection of non-vocabulary word in DTW and LPC based spoken word recognizer”, Journal ASI (NSA-2004), Vol 32, Mysore, Nov 25-27, 2004.

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

[8] Rabiner L, Juang B., “Fundamentals of Speech Recognition”, PHI Signal Processing Series, 1993.

[9] Alpay Koc, “Acoustic feature analysis for robust speech recognition”, MS Thesis in EE, Bogazici University, 2002

[10] Mobin Abdul, Agrawal S.S, “Role of filter-bank features in DTW based word recognition for saving the computational time and rejecting the non-vocabulary word”, Int. Conf. on Systemics Cybernetics and Informatics (SCI 2001,ISAS 2001), Volume XIII, 2001.

[11] Saleem M, Mobin Abdul, Mustafa K, “Optimization of input parameters for estimation of LP coefficients for isolated word recognition”, Intl Conf. on Systemic, Cybernetics and Informatics (ICSCI-2005) Hyderabad, pp. 391-393, Jan 06-09, 2005.