SPOKEN WORD RECOGNITION OF GERMAN DIGITS UTTERED BY NATIVE AND NON-NATIVE SPEAKERS

MOHD SALEEM
Research Scholar, Dept of Computer Science,
Jamia Millia Islamia,
New Delhi INDIA
mdsaleem@gmail.com

ABDUL MOBIN
Scientist, Central Electronics Engineering Research Institute, Delhi Centre, CSIR Complex,
New Delhi. INDIA

K. MUSTAFA
Reader, Dept of Computer Science, Jamia Millia Islamia, New Delhi
INDIA-25
kmfarooki@yahoo.com

IMRAN AHMAD
MS Student,
International School New Media, Luebeck
GERMANY
iahmad@isnm.de

ABSTRACT

The variability of speech between speakers is a hindrance for the robustness in automatic speech recognition (ASR). During recognition process, the lower accuracy has been reported due to the variance in pronunciation of the non-native speakers from those native speakers. In this paper we investigate the spoken word recognition of German digits for native (German) speakers and non-native (Indian) speakers. The experiments were carried out by exploiting two speech databases; one consists of voice samples from native speakers, the other one consisting of voice samples from non-native speakers. The comparison tests were performed for the recognition of speech samples from native and non-native speakers uttering the same German digits (null, eins, zwei,…., neun). The recognition accuracy achieved for native speakers were much higher than those achieved for non-native speakers. A progressive enhancement in the recognition accuracy has been achieved by clustering native and non-native voice samples in the training phase of the spoken word recognizer. Interpolating the acoustic model is also been tested for the enhancement in the recognition accuracy.

Keywords: Linear predictive coding, non-native speaker, Interpolation Model, Dynamic time warping.

1. INTRODUCTION

The urgent need of higher recognition accuracy for the non-native speech recognition systems comes into the picture due to global interaction between different countries and the peoples migrating from Indian origin to Germany are bound to learn the German language as the secondary languages. The increasing number of peoples learning German language leads to provide such a system which is feasible to both native and non-native speakers. Despite progressive enhancements from isolated word recognition to continuous speech recognition, the recognition accuracy has been observed to be drastically lower for non-native speakers due to the non-native accent and disfluent speech. The recognition accuracy achieved for non-native speakers is usually lower than that observed for the native speakers since the system were trained for native speakers [4]. The characteristic of non-native speech are very much different from the native speech. The syntactical errors, pronunciation errors and accented pronunciation are the characteristics of the non-native speech that depends on the speaker that how much he/she has learn the language. A higher acoustic variability is present in the non-native speech as compared to the native speech. Currently a number of procedures are being used for minimizing the effect in recognition accuracy for the non-native speech. The general approach is to use some non-native speech in the training phase with the native speech samples. Another approach is speaker adaptation techniques; some scientists also working on using multilingual HMM for non-native speech.

The spoken word recognizer for German digit is investigated by us. Two databases: one consists of voice samples from native speakers, the other one consisting of voice samples from non-native speakers were used [10]. The linear predictive coding is used to extract the features of voice samples and recognition experiments were performed for ten native (German) and ten non-native (Indian) speakers. The comparative study shows the lower accuracy for non-native speech. The system was trained for both native and non-native speech samples to improve the recognition performance. Another technique, interpolating the acoustic model is also been tested for the enhancement in the recognition accuracy [11][12].

2. 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 [5]. 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 [8]. 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. However, the 10-LP coefficients were computed from the wave files of spoken German digits by several speakers stored on a multimedia PC [9]. In this way a database consisting of 10 utterances of German digits (null through neun) spoken by ten native (German) and ten non-native (Indian) speakers were created and used for the recognition experiments [2][3].

3. INTERPOLATING ACOUSTIC MODEL

The basic recognition experiment for the native data does not give better result. The training of native data gives higher error rate as compared to the non-native data for the recognition of non-native speech data, and the combination of native German data with non-native data gives slightly better result [11][12]. The first experiment was performed by training the native speech data and non-native speech data in the ratio of two to one. We discussed here the interpolation of acoustic models for better representation of the native and non-native speech data to optimize the performance. While interpolating the both models, native models are better trained and the non-native data are more suitable for the test data [10]. Since we have two models to interpolate, the interpolation can be defined as:

$$I(V) = C_{Native}\, I_{Native}(V) + C_{Non\text{-}Native}\, I_{Non\text{-}Native}(V)$$

Where $C_{Native} + C_{Non\text{-}Native} = 1$;

V = Acoustic features Vector
$I(V)$ = Acoustic models;
C = Interpolation Constants

We have performed the number of experiments by varying the interpolation constants [4]. The interpolation constant 0 (zero) shows the performance with the native acoustic models, and the interpolation constant 1 (one) shows the performance with the non-native acoustic models. The figure in the experiment section shows the results at different interpolation constants.

4. SPOKEN WORD RECOGNIZER

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 $d(a_i,b_j) = a_i\text{-}b_j$ between two points a_i and b_j. The alignment of two time series samples can be found very efficiently 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:

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

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

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

The similarity between A and B is obtained as

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

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

$$W = \{ (i,j) \mid |i\text{-}j| \le r \} \qquad \ldots (4)$$

Where r is a positive integer, chosen so that the timing variation is practically limited within equation (4). Introduction of the restriction i.e. equation (4) reduces the amount of computation considerably [6].

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. The idea of the DTW technique is to match a test input represented by a multi-dimensional feature vector $A = [a_1,a_2,\ldots,a_I]$ with a reference template $B = [b_1,b_2,\ldots,b_J]$. 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 reference and test 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 speed up the distance computation by eliminating unlikely reference patterns, the cumulated distance rejection thresholds have been used. The accumulated distances using recurrence relation are sequentially computed starting from first LP coefficient to 10th LP coefficient followed by their successive additions (cumulative distance). If we denote the minimum cumulated distance as D_L, and the rejection threshold as T_L corresponding to L^{th} LP coefficient, then if

$$D_L > T_L \ (L=1,2,\ldots,12)$$

The computation is aborted at L^{th} LP coefficient and the test word corresponding to the reference pattern is rejected as a candidate for rejection. The methodology for the computation of T_L is based on the facts as follows. 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 A_L and B_L (reference and test) time-sampled speech patterns with respect to the output of the L^{th} LP coefficient [7].

5. EXPERIMENTS

The performance of the spoken word recognizer based on the interpolation of acoustic models was measured and compared with the baseline recognizer. The experiment 1 shows the performance of the baseline recognizer, while experiment 2 shows the performance of the recognizer where both the native and non-native speech data were used for the training. The experiment 3 shows the performance of the recognizer by interpolating the acoustic models for the native and non-native speech data. The word accuracy was computed and analyzed for spoken German digits (null through neun) for several native and non-native male speakers. The word accuracy (WA) is defined as

$$\text{WA(in \%)} = \frac{\text{No of correctly recognized word *100}}{\text{Total no of words in the test suite}}$$

The improvement in the recognition for baseline recognizer was achieved by giving little weight to the non-native speech data in the training phase of the recognizer. The 10 native speech utterance and 5 non-native speech utterance was used to train the speech recognizer (experiment 2), while interpolating acoustic models give a series of results by assigning the interpolation constant to the baseline native and baseline non-native models. The following table shows the word accuracy for the 10 speakers in the different experiments performed on the dynamic time warping and linear predictive coding based speech recognizer. Here the 100 tests were performed for the 10 male speakers uttering 10 spoken German digits.

Experiments	Word Accuracy (in %)	
	For Native	For Non-native
Experiment 1	93	56
Experiment 2	90	74
Experiment 3	92	86

Table: Word accuracy achieved in the different experiments

The following bar chart shows the performance of the recognizer for the different values of the interpolation constant of the acoustic models.

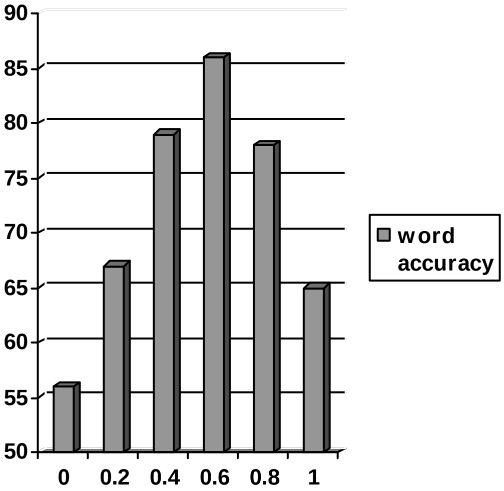

Figure: Performance of the spoken word recognizer for various interpolation constants

6. DISCUSSION AND CONCLUSION

In this paper we have investigated automatic recognition of German digits spoken by native speakers (German speakers) and non-native speakers (Indian speakers). We had explored how the acoustic models can be tailored for better recognition of non-native speech. As small amount of non-native speech data can be used very effectively for enhancement of recognition performance for non-native speech, using the interpolation constants we have given the weight to the model in the training phase of the spoken word recognizer. Significantly improved recognition accuracy has been achieved by interpolating the native and non-native models using interpolation constants.

As future work, we are planning to revamp the technique by using another feature extraction technique (Wavelet Transform) and also using larger number of database. Since the non-native speakers able to match only a fraction of pronunciation by native speakers, we also plan to evaluate, how much additional information about the native speech is efficient for robust recognition of non-native speech.

REFERENCES:

[1] Alpay Koc, "Acoustic feature analysis for robust speech recognition", MS Thesis in EE, Bogazici University, 2002

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

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

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

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

[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 pp. , Mysore, Nov 25-27, 2004.

[7] 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.

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

[9] Saleem M, Mobin Abdul, Mustafa K, "Optimization of input parameters for estimation of LP coefficients for isolated word recognition", proceeding Intl Conf. on Systemic, Cybernetics and Informatics (ICSCI-2005) Hyderabad, pp. 391-393, Vol. 2, Jan 06-09, 2005.

[10] Wang Z, Schultz T, "Non-native spontaneous speech recognition through polyphonic decision tree specialization", proc. EUROSPEECH, Geneva, pp. 1449-1452, 2003.

[11] Wang Z, Schultz T, Waibel A, "Comparison of acoustic model adaptation techniques on non-native speech", In proceeding of ICASSP, Vol-1, pp. 540-543, Hong Kong, April, 2003.

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