Research article Available online www.ijsrr.org ISSN: 2279-0543

International Journal of Scientific Research and Reviews

Pattern Normalization/Template Optimization in Order To Optimize Speech Recognition Process

Mutcha Srinivasa Rao*

Sr. Technical Officer & Project Leader. AAI, C-DAC
NSG IT Park, Aundh, Pune, Maharashtra, India – 411 007

ABSTRACT:

One of the main problems in speech recognition systems is the preparation of reliable reference templates for the set of words to be recognized. The accuracy of the speech recognition systems greatly relies on the quality of the prepared reference templates. The normal procedure in selecting the reference templates is to select one example then test its recognition rate. If the recognition rate is high then this reference is kept, otherwise another template has to be selected.
A common way to improve the recognition performance is to use several templates for each word. This procedure is computationally inefficient because it increases the number of templates. Vector quantization (VQ) is another solution to prepare reliable templates for the DTW-based speech recognition systems. However, it requires many training examples to prepare a reliable codebook. In order to keep up the computational efficiency feasible, a simple method is to use single reference templates per word. Even if single templates per word are chosen, there are two disadvantages:

- A single sample template per utterance cannot account for the variability of the speech signal. Even the same user cannot speak the same word exactly the same each time.
- There is no way to indicate the quality of information content contained in the sample template.

The solution to this problem is to make use of an Average Template Method that has the following features:

- Many samples are taken initially to account for variability in the speech signal for each word.
- It extracts the reference template from a set of examples rather than one example.
- The effect of a bad sample may be mitigated out of the good samples.
- No additional computational cost incurred during the overall recognition Process.

KEYWORDS: Pattern Normalization, Template Optimization, Speech Recognition, ASR, Speech Optimization, Dynamic Time Warping, Time Normalization.

Corresponding Author:-

Srinivasa Rao Mutcha
Sr. Technical Officer & Project Leader
Applied Artificial Intelligence Group (AAI),Centre for Development of Advanced Computing (C-DAC),NSG IT Park, Aundh, Pune, Maharashtra, India – 411 007
E-Mail: srinivas.mutcha@gmail.com, srinivasam@cdac.in

INTRODUCTION:

The pattern/template normalization technique using average template technique:

As mentioned earlier, the overall process of an Average Template technique involves using a number of samples initially to account for variability in the speech signal for each word. The overall definitive reference pattern/template is composed from a set of templates rather than one single template. This also serves as to improve the quality of the reference template/pattern by diminishing the effect of a low quality sample out of the overall sample cluster. The most crucial advantage is that the overall procedure is computationally efficient.

The process is depicted in the following figure:

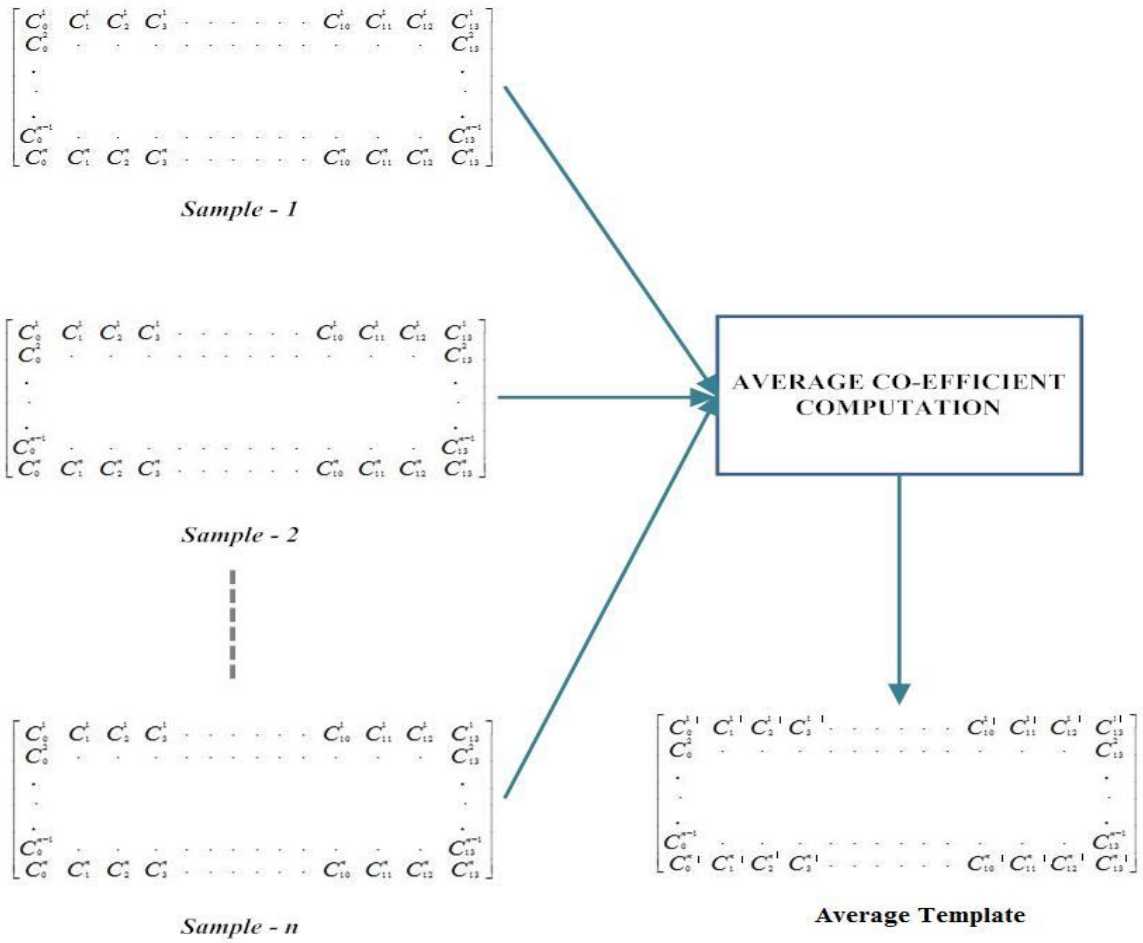

Fig.1 Average Template Computation

Briefly the process involves the following steps:

A. Choosing the Best Sample

- Calculate the average length of speech sample per each word.
- For each word, find the speech samples whose length is the nearest to the average length.

- That particular sample is chosen to be the best sample.

B. Time Normalization

- Adjust the length of other samples to be equal to the best sample by using a time-normalization technique

C. Average Template Creation

- Each Time-normalized sample is represented by a coefficient matrix.
- Average Template is created by averaging each coefficient with corresponding same row-column coefficient of all the other samples.
- The physical meaning of the average template is to create a template whose local magnitude spectrum is optimized over time and frequency dimensions.

TIME NORMALIZATION OF SAMPLES:

The traditional way of preparing the reference templates is by judiciously selecting one example for each word (needed to be recognized) and considering it as a reference template for that word. The disadvantage of using a single reference template is that it is not robust to the speech signal variability. That is because it is almost impossible for a person to repetitively speak a word exactly in the same way. The speech signal produced would vary according to many factors. Therefore, if the template created is bad, the user would have to change it until he finds a suitable template for the tested word. To overcome this problem without incurring more computations in the recognition phase, the average template technique is developed to prepare more robust templates, called crosswords reference templates (CWRTs). Using these templates can greatly improve the recognition accuracy, as it is prepared from multiple examples rather than just one example. However, the most important of the sample averaging technique discussed in the previous section involves Time Normalization of two different samples to make them equal in length. This can be achieved by using a Dynamic Programming technique known as Dynamic Time Warping or DTW.

DYNAMIC TIME WARPING:

The idea of the DTW technique is to match a test input represented by a multi-dimensional feature vector T = [t1, t2, t3,…..,tn]with a reference template R = [r1,r2,r3,…rm]

The following figure depicts graphically the idea of the DTW.

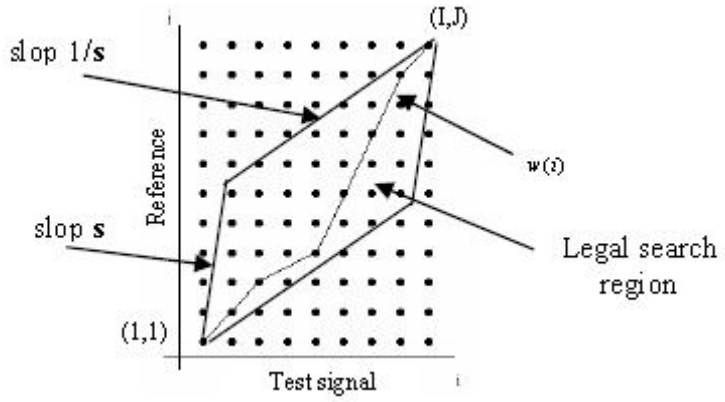

Fig. 2 Dynamic Time Warping for Template Optimization

The co-ordinate (i, j) is the location of the local distance between frame i to frame j. The aim of dynamic time warping is to find the function w(i) such that it gives the least cumulative difference between the compared signals.

The compress/expand algorithm to the initially selected reference template comprises of the following steps:

1. Align the first template with the initial reference template and find the optimum warping function w(i), $1 < i < m$.
2. Trace backward from the last frame to the first one, looking for the slope of the sub-paths for every frame of the speech signal along the alignment path w(i).
3. There are three slop possibilities to deal with:
    A. Slope is 1: Nothing is needed to be changed
    B. Slope is 2: The frame of the speech signal is replicated (expand), i.e., w(i-1) gets the identical frame to w(i).
    C. Slope is 0.5: An average frame is calculated from the two consecutive frames (compress), i.e., w(i) is merged with w(i-1) and gets their average value.
4. Repeat steps 1 and 2 for all the other templates of the available examples to get a set of equal length templates.
5. Average the aligned templates across each frame to get the final reference template.

CONCLUSION:

The reference template prepared by the mentioned average template technique is simply derived from a few examples of the words to be recognized rather than selecting just one example as a reference. It

differs from the VQ technique in a sense that it requires fewer examples, and it doesn't incur any quantization errors. In fact the VQ technique cannot work satisfactorily when only few training examples are available. The Reference Patterns/Templates thus created may be stored in the Acoustic Database.

REFERENCES:

[1] Robert Schalkoff: Pattern Recognition – Statistical, Structural and Neural Approaches. John wiley & sons, Inc. Ltd., Newyork.2007.

[2] Fukada T, Aveline S, Schuster M and Sagisaka Y. Segment boundary estimation using recurrent neural networks, Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997.

[3] Mutcha Srinivasa Rao, and ST Patil. An Automatic Speech Recognition System for Indian Language Computing using DTW, Proceedings of 2nd International Conference on Renewable Energy Resources and Management, Rajasthan, India. 2012.

[4] Unal FA and Tepedelenlioglu N. Dynamic time warping using an artificial neural network, volume 4, proceedings of International Joint Conference on Neural Networks, IJCNN -1992.

[5] Woojay Jeon, Changxue Ma. Efficient search of music pitch contours using wavelet transforms and segmented dynamic time warping, Proceedings of IEEE International conference on Acoustics, Speech and Signal Processing (ICASSP), 2011.

[6] Myers C, Rabiner L and Rosenberg A. Performance tradeoffs in dynamic time warping algorithms for isolated word recognition, Transactions of IEEE conference on Acoustics, Speech and Signal Processing, 2006.

[7] Guiling Li, Yuanzhen Wang, Min Li & Zongda Wu. Similarity Match in Time Series Streams under Dynamic Time Warping Distance, International Conference Computer Science and Software Engineering, 2008; 6:

[8] Fusheng Yu, Keqiang Dong, Fei Chen, Yongke Jiang and Wenyi Zeng. Clustering Time Series with Granular Dynamic Time Warping Method, IEEE International Conference on Granular Computing (GRC) 2007.

[9] Sajjan SC, Vijaya C. Comparison of DTW and HMM for isolated word recognition, International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME), 2012.

[10] Hsien Leing Tsai and Shie-Jue Lee. A neural network model for spoken word recognition, IEEE International Conference on Computational Cybernetics and Simulation, 1997.

[11] Rashwan MA, Fahmy MM. A new technique for speaker-independent isolated word recognition, International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1988.

[12] IEEE International Conference on "Acoustics, Speech, and Signal Processing (ICASSP), 1997.

[13] Shore J and Burton D. Discrete utterance speech recognition without time normalization, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 1982.

[14] Jong Gwan Lim, Sang-Youn Kim and Dong-Soo Kwon. Pattern recognition-based real-time end point detection specialized for accelerometer signal, ASME International Conference on Advanced Intelligent Mechatronics (AIM) 2009.