

**Enhanced Techniques to Improve Speech Query based  
Tamil to English Cross-Language Text Retrieval System  
using Ensemble Models**

*Thesis submitted in Partial Fulfilment of the*

**Degree of Doctor of Philosophy (Ph.D.)**

*By*

**P.Iswarya**

**11PH34**

**Department of Computer Science**

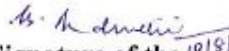
*Avinashilingam Institute for Home Science and Higher Education for Women,*

*Coimbatore – 641043*

**August 2015**

## CERTIFICATE

I certify that the thesis entitled “Enhanced Techniques to Improve Speech Query based Tamil to English Cross-Language Text Retrieval System using Ensemble Models ”submitted for the degree of Doctor of Philosophy (Ph.D.) by Mrs. P.Iswarya, is the record of research work carried out by her during the period from June 2011 to August 2015 under my guidance and supervision, and that this work has not formed the basis for the award of any Degree, Diploma, Associateship, Fellowship or other Titles in this University or any other University or Institution of Higher Learning.

  
Signature of the <sup>10/8/15</sup>

**Head of the Department**

**Dr. G. PADMAVATHI**  
M.Sc., M.Phil., Ph.D.  
Professor and Head  
Department of Computer Science  
Avinashilingam Institute for Home Science  
and Higher Education for Women  
Coimbatore - 641 043

  
Signature of the Supervisor

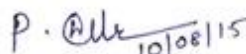
**Dr. (Mrs) V. RADHA**  
Professor, Dept. of Computer Science  
Avinashilingam Institute for Home Science  
and Higher Education for Women  
Coimbatore - 641 043

  
Signature of the Dean <sup>10-08-15</sup>

Dean, Faculty of Home Science  
Avinashilingam Institute for Home Science  
and Higher Education for Women  
Avinashilingam University  
Coimbatore-641 043

## DECLARATION

I declare that the thesis entitled “**Enhanced Techniques to Improve Speech Query based Tamil to English Cross-Language Text Retrieval System using Ensemble Models** ” submitted by me for the degree of **Doctor of Philosophy (Ph.D.)** is the record of work carried out by me during the period from June 2011 to August 2015 under the guidance of **Dr. (Mrs.) V.Radha**, Professor, Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, and has not formed the basis for the award of any Degree, Diploma, Associateship, Fellowship, Titles in this University or any other University or other similar Institution of Higher Learning.

  
Signature of the Candidate

## ACKNOWLEDGEMENT

*First and Foremost I would like to Thank **God**. You have given me the power to believe in myself and pursue my dreams. I could never have done this without the faith I have in you, the Almighty.*

*I would like to thank the Chancellor **Dr. T.S.K. MEENAKSHI SUNDARAM**, M.A., M.Phil., Ph.D., Avinashilingam Institute for Home Science and Higher Education for Women, for providing an opportunity to conduct the study.*

*I express my heartfelt gratitude to **Dr. (Mrs.) SHEELA RAMACHANDRAN**, M.Sc., P.G. Dip., Ph.D., Vice Chancellor, Avinashilingam Institute for Home Science and Higher Education for Women, for her supportive attitude and encouragement.*

*I offer my profound gratitude to **Dr. (Mrs.) A.VENMATHI**, M.Sc., Dip.Ed., M.Phil., Ph.D., Registrar , Avinashilingam Institute for Home Science and Higher Education for Women, for the Administrative Support rendered throughout the span of the research work.*

*I express my sincere thanks to **Dr. (Mrs.) A.PARVATHI**, M.Sc., Dip.Ed., M.Phil., Ph.D., Dean, Faculty of Science , Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for her spontaneous and timely help and amenities provided for the successful completion of this research work.*

*I also extend my thanks to **Dr. (Mrs) G. PADMAVATHI**, M.Sc., M.Phil., Ph.D., Professor and Head, Department of Computer Science , Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for her support, encouragement and cooperation rendered towards the completion of this research.*

*I take immense pleasure to express my sincere and deep sense of gratitude to my Supervisor **Dr. (Mrs). V. RADHA**, M.Sc., P.G.D.C.A., P.G.D.O.R., B.Ed., M.Phil., Ph.D., Professor, Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for her sustained enthusiasm, creative suggestions, motivation and exemplary guidance throughout the course of my doctoral research. Apart from the subject of my research, I learnt a lot from him, which I am sure, will be useful in different stages of my life. I solemnly submit my honest and humble thanks to her for bringing my dreams into reality.*

*I am very much grateful to **Mr. KARNAMURTHI** for correcting all the grammatical mistakes in the write up. I deeply appreciate their timely help and constructive criticism which brought the document to shape.*

*I accord my warm thanks to all the **MEMBERS OF THE FACULTY, NON-TEACHING STAFF, AND RESEARCH SCHOLARS** from the Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for their cooperation.*

*A special thanks to my family. Words cannot express how grateful I am to my **PARENTS** for all of the sacrifices that you've made on my behalf. Your prayer for me was what sustained me thus far. I would also like to thank all of my **FRIENDS** who supported me in writing, and incited me to strive towards my goal. At the end I would like express appreciation to my beloved husband **Mr. R. GANESAN** who spent sleepless nights with and was always my support in the moments when there was no one to answer my queries.*

**P. ISWARYA**

# CONTENTS

<b>CHAPTER NO</b>	<b>TITLE</b>	<b>PAGE NO</b>
	<b>LIST OF TABLES</b>	
	<b>LIST OF FIGURES</b>	
	<b>LIST OF ABBREVIATIONS</b>	
	<b>ABSTRACT</b>	
<b>1</b>	<b>INTRODUCTION</b>	<b>1</b>
<b>1.1</b>	Information Retrieval	<b>1</b>
<b>1.2</b>	Cross-Language Information Retrieval	<b>4</b>
<b>1.3</b>	Types of CLIR Systems	<b>6</b>
<b>1.4</b>	Importance of CLIR in Tamil-English Text Retrieval	<b>6</b>
<b>1.5</b>	Challenges in CLIR with Tamil Language	<b>8</b>
<b>1.6</b>	Overview of Techniques used in Proposed CLIR System	<b>11</b>
<b>1.6.1</b>	Speech Query and Recognition	<b>11</b>
<b>1.6.2</b>	Query Translation	<b>13</b>
<b>1.6.3</b>	Document Retrieval	<b>15</b>
<b>1.7</b>	Motivation and Objectives	<b>16</b>
<b>1.8</b>	Layout of the Chapters	<b>18</b>
<b>1.9</b>	Chapter Summary	<b>19</b>
<b>2</b>	<b>REVIEW OF LITERATURE</b>	<b>20</b>
<b>2.1</b>	Tamil Language	<b>20</b>
<b>2.1.1</b>	Tamil Script	<b>21</b>
<b>2.2</b>	Speech Recognition	<b>25</b>
<b>2.2.1</b>	Historical background of Speech Recognition	<b>25</b>
<b>2.2.2</b>	License Free Speech Recognition Software	<b>28</b>

<b>CHAPTER NO</b>	<b>TITLE</b>	<b>PAGE NO</b>
2.2.3	Commercial Speech Recognition Software	28
2.2.4	Speech Recognition for Indian Languages	29
2.2.5	Tamil Speech Recognition System	33
2.3	Information Retrieval (IR) Approaches	35
2.3.1	Boolean Model	35
2.3.2	The Vector Space Model	36
2.3.3	Probabilistic Retrieval Model	37
2.3.4	The Language Model	39
2.3.5	Syntactic Models	41
2.4	Cross-Language Information Retrieval	42
2.4.1	History of CLIR	42
2.4.2	Tamil-based CLIR Systems	46
2.5	Chapter Summary	49
<b>3</b>	<b>METHODOLOGY</b>	<b>51</b>
3.1	Proposed Methodology	51
3.2	Phase I : Speech Recognition	54
3.3	Phase II : Query Translation	57
3.4	Phase III :Text Retrieval	61
3.5	Performance Evaluation	63
3.6	Chapter Summary	64
<b>4</b>	<b>DESIGN OF TAMIL SPEECH QUERY RECOGNITION SYSTEM</b>	<b>65</b>
4.1	Pre-processing	67
4.1.1	Pre-Emphasis	68
4.1.2	Noise Removal Algorithm	68
4.1.3	Windowing	78
4.1.4	Silence Removal	79
4.2	Feature Extraction	81
4.2.1	Wavelet Packet Decomposition	83
4.2.2	Conventional MFCC Feature Extraction	84

<b>CHAPTER NO</b>	<b>TITLE</b>	<b>PAGE NO</b>
4.2.3	Proposed Wavelet-Packet-Based MFCC Feature Extraction Technique	86
4.2.4	Conventional LPCC Feature Extraction	88
4.2.5	Proposed Wavelet-Packet-Based LPCC Feature Extraction Technique	90
4.3	Speech Query Recognition	91
4.3.1	Support Vector Machine	91
4.3.2	Ensemble SVM-Based Classifier	95
4.4	Chapter Summary	99
<b>5</b>	<b>DESIGN OF QUERY TRANSLATION SYSTEM</b>	<b>100</b>
5.1	Tokenization	101
5.2	Pre-processing	102
5.2.1	POS Tagging	102
5.2.2	Chunking	115
5.2.3	Named Entity Recognition (NER)	115
5.2.4	Morphological Analysis	115
5.2.5	Word Sense Disambiguation	119
5.3	Translation	123
5.4	Transliteration with Error Correction	125
5.5	Chapter Summary	125
<b>6</b>	<b>DESIGN OF TEXT RETRIEVAL SYSTEM</b>	<b>126</b>
6.1	Query Expansion	127
6.2	Document Retrieval	129
6.2.1	Improved KNN Classification	129
6.2.2	Improved ARC Algorithm	136
6.2.3	Ensemble Model based on Hybrid Classifier	144
6.3	Ranking	146
6.4	Chapter Summary	147

<b>CHAPTER NO</b>	<b>TITLE</b>	<b>PAGE NO</b>
<b>7</b>	<b>RESULTS AND DISCUSSION</b>	<b>148</b>
7.1	Experimental Setup	149
7.1.1	Datasets Used	149
7.1.2	Performance Metrics	150
7.2	Speech Recognition	155
7.2.1	Pre-processing	156
7.2.2	Feature Extraction and Recognition	161
7.3	Query Translation	166
7.3.1	POS tagging	167
7.3.2	Morphological Analysis	170
7.3.3	Word Sense Disambiguation	172
7.3.4	Translation and Transliteration	175
7.4	Text Retrieval	178
7.5	Visual Results	188
7.5.1	Query1 Visual Results	188
7.5.2	Query2 Visual Results	191
7.6	Chapter Summary	194
<b>8</b>	<b>SUMMARY AND CONCLUSION</b>	<b>195</b>
	<b>BIBLIOGRAPHY</b>	<b>200</b>
	<b>PUBLICATIONS RELATED TO RESEARCH WORK</b>	<b>226</b>

## LIST OF TABLES

TABLE NO	TITLE	PAGE NO
2.1	Tamil Scripts	22
2.2	CLIR Experiments in TREC	44
5.1	POS Tagset	103
5.2	Description of Features for Tamil POS Tagging	106
5.3	Case Suffixes used with Noun	116
5.4	Case Suffixes used with Verb	117
5.5	Sample entries in Sense-Collocation Dictionary	123
7.1	Coding Scheme used in Phase I	155
7.2	Sentence Level Accuracy (%) of ATSQR System While Using the Top 20 Tamil Queries in FIRE Dataset	162
7.3	Coding Scheme used in Phase II	166
7.4	Precision, Recall and F-Measure of POSSA and POSHES Algorithms	167
7.5	POS Tagging Trials and Errors	170
7.6	Performance of Morphological Analyzer Precision (%)	171
7.7	Performance of Morphological Analyzer Recall (%)	171
7.8	Performance of Morphological Analyzer F-Measure (%)	171
7.9	Performance of Morphological Analyzer Accuracy (%)	172
7.10	Performance of Morphological Analyzer Speed (Seconds)	172
7.11	Average Performance of WSD Algorithm	173
7.12	Performance of WSD Algorithms	174
7.13	F-Measure of the Query Translation Algorithms	175
7.14	Coding Scheme used in Phase III	179
7.15	F-Measure of Text Retrieval Systems with Speech-Based STTQ	179
7.16	F-Measure of Text Retrieval Systems with Text-Based STTQ	180
7.17	F-Measure of Text Retrieval Systems with Text-Based DTQ	180
7.18	F-Measure of Text Retrieval Systems with Text-Based NTQ	181
7.19	Evaluation of the Mono-lingual and Cross-Lingual Runs	186
7.20	Speed (Seconds) of CLTR Systems	187

## LIST OF FIGURES

FIGURE NO	TITLE	PAGE NO
1.1	Processes of IR Systems	3
1.2	CLIR Scenario	4
1.3	Processes in CLIR System	5
1.4	ASR System	12
1.5	Query Translation Techniques	16
2.1	Combinant Tamil Letters	23
3.1	Steps in Proposed CLTR System	53
3.2	Research Methodology	55
4.1	Steps in Proposed ATSQR System	67
4.2	Speech Waveform of Word “amma (அம்மா)” before and after pre-emphasis filter	69
4.3	W2SHT Algorithm	70
4.4	DWT Signal Analysis	71
4.5	D4 Forward Transformation	74
4.6	Effect of Thresholding	75
4.7	D4 Inverse Transformation	78
4.8	Overlapping Windowing Scheme	79
4.9	Speech and Silence Identification	82
4.10	3-Level Wavelet Packet Decomposition	84
4.11	MFCC Feature Extraction Process	84
4.12	WPMFCC Feature Extraction Process	86
4.13	Block diagram of LPCC Feature Extraction	88
4.14	Steps in WPLPCC Feature Extraction Technique	90
4.15	Support Vector Machine Hyperplane	93
4.16	Boosting Algorithm	97
5.1	Architecture of TQTS	101
5.2	Tokenization Pseudo-code	102
5.3	Ensemble Feature Selection	109

<b>FIGURE NO</b>	<b>TITLE</b>	<b>PAGE NO</b>
5.4	Architecture of WNN	110
5.5	Support Vector Machine Hyperplane	113
5.6	Hybrid SVM-WNN Classifier	114
5.7	General Framework of Morphological Analyzer	116
5.8	Structure of Pronoun Word Form	118
5.9	Process of Disambiguating Word Sense using POS tagging	121
5.10	Architecture of WSD using Clustering and Sense-Collocation Dictionary	122
5.11	Representation of Context Space	122
5.12	Tense Marker and Gerund Ending Rules	124
6.1	General Architecture of a Document Retrieval System	126
6.2	QE Algorithm	129
6.3	Improved KNN Classification Algorithm	131
6.4	Single Pass Clustering Algorithm	133
6.5	Automatic Threshold Estimation Procedure	134
6.6	KNN Classification	135
6.7	Steps in IARC Algorithm	137
6.8	Basic Association Mining Algorithm	139
6.9	Apriori Algorithm	140
6.10	ARC-BC Algorithm	141
6.11	Classification Algorithm	142
6.12	Database Coverage	143
6.13	Proposed EHAKNN Classifier	144
6.14	HAKNN Classifier	145
7.1	FIRE 2011	151
7.2	Signal used for Calculating MSE	152
7.3	Average SNR Analysis of White Noise Removal	157
7.4	Average SNR Analysis of Babble Noise Removal	157
7.5	Average SNR Analysis of External Noise Removal	158
7.6	Average MSE Analysis of White Noise Removal	158

<b>FIGURE NO</b>	<b>TITLE</b>	<b>PAGE NO</b>
7.7	Average MSE Analysis of Babble Noise Removal	159
7.8	Average MSE Analysis of External Noise Removal	159
7.9	Average Speed (Seconds) of Pre-processing Algorithms	161
7.10	Effect of Pre-processing and Feature Extraction Techniques with different classifiers on ATSQR with respect to Average Accuracy (%)	164
7.11	Speed (Seconds) of Feature Extraction and Recognition	165
7.12	Accuracy of POS Tagging Algorithms	169
7.13	Speed of POS Tagging Algorithms	169
7.14	Query Translation using STTQ	176
7.15	Query Translation using DTQ	177
7.16	Query Translation using NTQ	177
7.17	Speed (Seconds) of the Query Translation Algorithms	178
7.18	Macro F-Measure of Text Retrieval Systems with Speech-Based and Text-Based STTQ	181
7.19	Macro F-Measure of Text Retrieval Systems with Text-Based with DTQ	182
7.20	Macro F-Measure of Text Retrieval Systems with Text-Based NTQ	182
7.21	Average Accuracy of the Text Retrieval Systems with STTQ	183
7.22	Average Accuracy of the Text Retrieval Systems with DTQ	184
7.23	Average Accuracy of the Text Retrieval Systems with NTQ	184
7.24	P-R Curve for CLTR Systems using STTQ Queries	185
7.25	P-R Curve for CLTR Systems using DTQ Queries	185
7.26	P-R Curve for CLTR Systems using NTQ Queries	186
7.27	Original Noisy Query 1	188
7.28	Query 1 After Noise Removal	188
7.29	Query 1 After Silence Removal	189
7.30	Query 1 Speech Recognition	189

<b>FIGURE NO</b>	<b>TITLE</b>	<b>PAGE NO</b>
<b>7.31</b>	Query 1 Morphological Analysis	<b>190</b>
<b>7.32</b>	Input Query1 to TECLTR-T	<b>190</b>
<b>7.33</b>	Documents Retrieved for Query1 (Swine Flu Vaccine)	<b>191</b>
<b>7.34</b>	Original Noisy Query 2	<b>191</b>
<b>7.35</b>	Query 2 After Noise Removal	<b>191</b>
<b>7.36</b>	Query 2 After Silence Removal	<b>192</b>
<b>7.37</b>	Query 2 Speech Recognition	<b>192</b>
<b>7.38</b>	Query 2 Morphological Analysis	<b>193</b>
<b>7.39</b>	Input Query2 to TECLTR-T	<b>193</b>
<b>7.40</b>	Documents Retrieved for Query2 (Price Hike of Petroleum Products)	<b>194</b>

---

## LIST OF ABBREVIATIONS

---

<b>ADABOOST</b>	ADaptive BOOSTing
<b>AMOS</b>	Archived Multi-Objective Simulated Annealing
<b>ANN</b>	Artificial Neural Network
<b>AP</b>	Average Precision
<b>ARC</b>	Associative Rule-based Classifier
<b>ARC-BC</b>	Association Rule-based Categorizer- By Category
<b>ARM</b>	Association Rule Mining
<b>ASR</b>	Automatic Speech Recognition
<b>ASRS</b>	Automatic Speech Recognition System
<b>ATSQR</b>	Automatic Tamil Speech Query Recognition
<b>BIM</b>	Binary Independence Model
<b>BM</b>	Best Match
<b>CLEF</b>	Cross-Language Evaluation Forum
<b>CLIR</b>	Cross-Language Information Retrieval
<b>CLTR</b>	Cross-Language Text Retrieval
<b>CS</b>	Cluster Score
<b>CSL</b>	Computer Speech Laboratory
<b>CSR</b>	Continuous Speech Recognition
<b>CWFT</b>	Context Words Features of a Token
<b>CWSD</b>	Conventional Word Sense Disambiguation Algorithm
<b>D4</b>	Daubechies wavelet-4
<b>DCM</b>	Dirichlet Compound Multinomial
<b>DCT</b>	Discrete Cosine Transformation
<b>DFR</b>	Deviation From Randomness
<b>DFT</b>	Discrete Fourier Transform
<b>DT</b>	Decision Tree
<b>DTQ</b>	Descriptive Topic Queries

---

---

## LIST OF ABBREVIATIONS

---

<b>DTW</b>	Dynamic Time Wrapping
<b>DWPT</b>	Discrete Wavelet Packet Transformation
<b>DWT</b>	Discrete Wavelet Transform
<b>EBMT</b>	Example Based Machine Translation
<b>ECLTR</b>	Existing Cross-Language Text Retrieval
<b>EFS</b>	Ensemble Feature Selection
<b>EHAkNN</b>	Ensemble Hybrid ARC and kNN Classifier
<b>ESMA</b>	Ensemble SVM-based Morphological Analyzer
<b>FIRE</b>	Forum for Information Retrieval and Evaluation
<b>FIRE</b>	Forum for Information Retrieval Evaluation
<b>HAkNN</b>	Hybrid Association Rule Mining and Improved kNN
<b>HLDA</b>	Heteroscedastic Linear Discriminant Analysis
<b>HMM</b>	Hidden Markov Model
<b>HQT</b>	Hybrid Query Translation
<b>HTK</b>	Hidden markov model Tool Kit
<b>IARC</b>	Improved Associative Rule-based Classifier
<b>IDWT</b>	Inverse Discrete Wavelet Transform
<b>IkNN</b>	Improved kNN
<b>ILIR</b>	Indian Language Information Retrieval
<b>IR</b>	Information Retrieval
<b>IT</b>	Information Technology
<b>IWR</b>	Isolated Word Recognition
<b>KL</b>	Kullback-Leibler
<b>kNN</b>	K-Nearest Neighbor
<b>LDA</b>	Linear Discriminant Analysis
<b>LPCC</b>	Linear Predictive Cepstral Co-efficient
<b>LVCSR</b>	Large Vocabulary Continuous Speech Recognition

---

---

## LIST OF ABBREVIATIONS

---

<b>MA</b>	Morphological Analyzer
<b>MAD</b>	Median Absolute Deviation
<b>MAP</b>	Mean Average Precision
<b>MAP</b>	Mean Average Precision
<b>MFCC</b>	Mel Frequency Cepstral Co-efficients
<b>MFR</b>	Multiple Frame Rate
<b>MFS</b>	Multiple Frame Size
<b>MLLR</b>	Maximum Likelihood Linear Regression
<b>MLP</b>	Multi-Layer Perceptron
<b>MLTR</b>	Mono-Lingual Text Retrieval
<b>MOO</b>	Multi-Objective Optimization
<b>MSE</b>	Mean Square Error
<b>MSE</b>	Mean Square Error
<b>MT</b>	Machine Translation
<b>NER</b>	Named Entity Recognition
<b>NICO</b>	Neural Inference COmputation
<b>NII</b>	National Institute for Informatics
<b>NIST</b>	National Institute of Standards and Technology
<b>NLP</b>	Natural Language Processing
<b>NTQ</b>	Narrative Topic Queries
<b>OOV</b>	Out-of-Vocabulary
<b>PCA</b>	Principal Component Analysis
<b>POS</b>	Part-of-Speech
<b>POSHES</b>	POS Tagging using Hybrid SVM-WNN Ensemble Classifier with Ensemble Feature Selection

---

---

## LIST OF ABBREVIATIONS

---

<b>POSSA</b>	POS Tagging using SVM classifier with Simulated Annealing and AMOSA
<b>PRF</b>	Pseudo-Relevance Feedback
<b>PRP</b>	Probability Ranking Principle
<b>QE</b>	Query expansion
<b>QT</b>	Query Translation
<b>S</b>	Conventional SVM Classifier
<b>SA</b>	Simulated Annealing
<b>SCP</b>	Support-Confidence based algorithm
<b>SDK</b>	Software Development Kit
<b>SDR</b>	Spoken Document Retrieval
<b>SES</b>	Ensemble SVM Classifier enhanced with SOM
<b>SMA</b>	SVM-based Morphological Analyzer
<b>SNR</b>	Signal to Noise Ratio
<b>SNR</b>	Signal to Noise Ratio
<b>SOM</b>	Self Organizing Map
<b>SRDSTQ</b>	Silence Removal from Denoised Speech Tamil Query
<b>SS</b>	SVM classifier enhanced with SOM
<b>STE</b>	Short Time Energy
<b>STTQ</b>	Short Title Topic Queries
<b>SVC</b>	Support Vector Classification
<b>SVM</b>	Support Vector Machine
<b>SVR</b>	Support Vector Regression
<b>TDIL</b>	Technology Development for Indian Languages
<b>TECLTR-S</b>	Tamil-English Cross-Language Text Retrieval for Speech query

---

---

## LIST OF ABBREVIATIONS

---

<b>TECLTR-T</b>	Tamil-English Cross-Language Text Retrieval for Text query
<b>TF-IDF</b>	Term Frequency-Inverse Document Frequency
<b>TQTS</b>	Tamil Query Translation System
<b>TQTS</b>	Tamil Query Translation System
<b>TREC</b>	Text Retrieval Conference
<b>VSM</b>	Vector Space Model
<b>W2SHT</b>	Wavelet Denoising based on Switching Soft and Hard Thresholding
<b>WHT</b>	Wavelet Denoising based on Hard Thresholding
<b>WNN</b>	Wavelet Neural Network
<b>WOPP</b>	Without Pre-Processing
<b>WPD</b>	Wavelet Packet Decomposition
<b>WPLPCC</b>	Wavelet Packet Based LPCC
<b>WPMFCC</b>	Wavelet Packet Based MFCC
<b>WPP</b>	With Pre-Processing
<b>WSD</b>	Word Sense Disambiguation
<b>WSDCS</b>	Word Sense Disambiguation Algorithm enhanced with Clustering and Sense-collocation Dictionary
<b>WSDPCS</b>	Enhanced WSD with Part-of-speech and Clustering based Sense-collocation
<b>WST</b>	Wavelet Denoising based on Soft Thresholding
<b>WWQT</b>	Word by Word Query Translation
<b>WWW</b>	World Wide Web
<b>ZCR</b>	Zero Crossing Rate

---

## **ABSTRACT**

Identification of techniques to improve the process of searching and retrieving the interested documents from large databases is a challenging task in information retrieval systems. Cross-Language Information Retrieval (CLIR), where queries and documents to be retrieved are in different languages, has become one of the major topics within the world of information retrieval community.

The main objective of this research work is to design and develop Tamil-English Cross-Language Text Retrieval (CLTR) systems based on text or speech query, by integrating key technologies like speech analysis, translation, document retrieval and ranking along with ensemble machine learning for accurate relevant English document retrieval. To achieve this objective, the research methodology was designed in three phases, where the first phase focused on Tamil query speech recognition, second phase focused on query translation (from Tamil to English) and the third phase concentrated on relevant English documents.

Speech recognition was performed in three steps. The first step performed noise removal, using wavelet-based denoising algorithm using Switching Soft and Hard Thresholding and silence removal based on two criteria, short time energy and zero crossing rate combined with automatic threshold estimation and rule-based algorithms. Feature extraction was performed using wavelet packet based Mel Frequency Cepstral Co-efficients and Linear Predictive Cepstral Co-efficients, which was standardized to same length using Self Organizing Map. The speech recognition is performed using an ensemble SVM classifier.

The query translation was performed using rule-based machine translation and statistical machine translation. Rule-based machine translation includes tokenization, Part-of-Speech (POS) tagging, chunking, morphological analysis, named entity recognition, word sense disambiguation and translation. The statistical machine translation is applied to carry out transliteration procedure.

The proposed POS tagging algorithm performs feature extraction and selections using three feature selection algorithms, whose output are combined to form an optimal

feature set. This is used by a fast SVM classifier to produce an optimized training vector that trains WNN classifier to improve the accuracy of POS tagging. The morphological analysis was performed using an ensemble SVM classifier with trained models. The model segments the words into morphemes, and it is tagged using corresponding labels. The word sense disambiguation was performed using an ensemble K-Means clustering algorithm with sense-collocation dictionary. The root words are translated using knowledge sources, and grammatical categories are replaced by implementing tense marker and gerund ending rules. Untranslated named entities are transliterated using statistical machine translation. Finally rule-based re-ordering is implemented to get correct structure of English sentence.

The final phase of the research work performs document retrieval using an ensemble Hybrid classification system using improved K-Nearest Neighbor classifier enhanced through the use of associative rule mining. The ranking of the retrieved documents was done using Okapi BM25 algorithm.

Experiments to evaluate the performance of the proposed algorithms were performed using Tamil queries from Forum for Information Retrieval and Evaluation (FIRE) dataset 2011. Three types of queries namely, short term title queries; descriptive title queries and narrative title queries were used. Performance metrics like Signal to noise ratio, Mean Square error, Speech recognition accuracy and speed were used to evaluate algorithms in Phase I. Phase II algorithms were evaluated using Precision, Recall, F- Measure, Accuracy and Speed. Phase III algorithms were analyzed using Precision, Recall, Mean Average Precision (MAP), F-Measure, Accuracy and Speed of retrieval.

Experimental results revealed that the proposed Tamil-English Cross -Language Text Retrieval with speech-based and text-based queries produced improved results, when compared with the conventional and existing algorithms, while using the various proposed algorithms in each phase.