ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

A Dynamic Technique for Speech Optimization and Speech-To-Speech Linguistic Translation

Protik Ganguly

Former B.E Student, Dept. of Information Technology, Mumbai University, India

ABSTRACT: Speech recognition and language translation are the emerging technologies in the field of Computer Science and Artificial Intelligence. Such technologies have given a new dimension to the world of communication, thus, globalization is getting unstoppable and languages are no more a barrier.This paper describes a novel approach to achieve high accuracy in speech recognition, speech-to-speech translation, machine translation and speech synthesis. The conventional methods are a text-to-text translation among different languages; which is pretty much like a dictionary. The proposed technique in this paper is divided into three modules – speech-to-text, text-to-text (translation) and text-to-speech, so to maintain high accuracy as a whole, accuracy stays uncompromised at every level/module.

KEYWORDS: Speech-to-speech; text-to-text; speech recognition; speech synthesis; machine translation

I. Introduction

Speech-to-speech translation is a technique for translating speech in one language to another language. In this paper,the whole process is discussed and dissected into three modules namely:i) speech-to-text, ii) text-to-text (actual language translation) and iii) text-to-speech.

i) Speech-to-Text

This module is based on speech recognition for converting the speech into text. The spoken words or phrase by the speaker gets recorded by the microphone and is further analysed by the system for conversion to text. This system is also known as "automatic speech recognition", "ASR", "computer speech recognition", "speech to text", or just "STT".

This process is divided into two parts, begins when speaker speaks words and sentences through the microphone. The speech is turned into a waveform using acoustic modeling (Acoustic modeling is used to create statistical representations of the sounds that make up each word. It is used by a speech recognition engine to recognize speech) by the system which undergoes preprocessing called speech recognition. The system determines the pauses and the extraneous sounds and picks up the actual speech. Then the speech signal is converted into a sequence of vectors, the words we speak are transformed into digital forms of the basic speech elements (phonemes) which are measured throughout the duration of the speech signal.

Once the first part of the process is over language modeling begins wherein the system actually begins to work. The language is compared to the digital dictionary that is stored in computer memory. This is a large collection of words, usually more than 100,000. When it finds a match based on the digital form it displays the words on the screen. This is the basic process for all speech recognition systems and software.

ii) Text-to-Text

This is the main module as it will solve the purpose of the whole study – the actual language translation called Machine Translation. The text that is developed from the speech is converted into the desired language by getting parsed through all the language rules. So to convert the source language into target language the system has to undergo two processes namely:

- Decoding the meaning of the source text
- Re-encoding the meaning of source text into the target text

DOI: 10.15680/IJIRCCE.2016. 0405182

Although, in the simplest of explanation decoding the meaning of source text may seem like a cakewalk but in reality it is an extremely complex cognitive process. So to decode the meaning of the source text in its entirety the system needs to interpret and analyse all the features of the source text for which it needs to have in-depth knowledge of grammar, semantics, syntax, and idioms etc. for smooth conversion. Similarly, that translator needs to have same kind of knowledge of the target text for Re-encoding. There are multiple approaches to implement this module of machine translation for example: Dictionary-based, Interlingual, Rule based and so on. But to keep the accuracy high and size of Text corpus low, for smooth and smart conversion we would review this module using hybrid approach(stastistical and example based).

iii) Text-to-Speech

In this module a technique called speech synthesis will be implemented, which is used to artificially produce human voice. The main aim of this module is to convert normal language text into speech. Synthesized speech is usually produced by concatenating pieces of recorded speech that are stored in database or corpus. Conventional systems store phonemes that provide largest output range but quality gets compromised. So we would use corpus based speech synthesis so that the output range remains large and high clarity is also achieved. Corpus Method is very popular for its high quality and natural speech output. The basic idea of corpus based speech synthesis or unit selection is to use the entire speech corpus as the audio inventory and to select the run time from this corpus, the longest available strings of phonetic segments that match a sequence of target speech sounds in the utterance to be synthesized thereby minimizing the number of concatenations and reducing the need for signal processing. One main drawback of Corpus method is relative weighing of acoustic distance measures.It needs large speech database with optimal coverage of target domain which is often the whole language. But this gives the best result when a highly flexible system is needed.

II. Related work

The earlier basic systems of machine translation were based upon rule based approach. Rule based systems used linguistic information like morphological, syntactic, and semantic information from dictionaries and grammars of both source and target languages for translation.Few systems like Systran, Japanese MT systems are the ones that are based on this approach and are simple in nature with advantages like no requirement of bilingual texts, domain independent, total control as rules were handwritten. But RBMT faced a drawback due to insufficient amount of good dictionaries and building new dictionary is expensive, some linguistic information still needs to be set manually and it failed to adapt with new domains. So the urge of new approach arose and then the example based approach came into the scenario.

In [1] author has reviewedExample based approach for machine translation. He explained how the size of corpus can be reduced by using example based approach. Traditionally, system used large size of text corpora for conversion which caused delay in final result and the clarity of conversion was not very delightful. But overtime machine translation has got sophisticated. In example based machine translation the system acts like a human being it goes through translated documents to understand the texts. Then on the basis of those sample documents it would use its artificial intelligence and translate further documents. This approach although reduced the size of corpus and human dependency yet, accuracy remained an issue as the system’s understand of language rules gets constrained to those documents only.

A system called Pangloss is developed by the author in [2]. The Pangloss system is implemented using example based approach. It uses no prior knowledge of language but only corpus, dictionary and few sentences. This system showed great error smoothening curves but its drawback is its dependency on the dictionary and the documents it refers for translation as the quality of these documents and dictionary is not up to the mark. Yet, the strength of this system lies in its sophisticated degradation which occurs only when there is discrepancy between the example corpus and the sentence to be translated. In [3], the author clearly explicated the drawbacks of EBMT. One issue is similar to that of the Pangloss system mentioned in [2] that usually there is discrepancy between sentence to be translated and the example fragment. Then there is a delay in retrieving the example for translation from the database and lastly, exploiting the retrieved translated example to actually translating the sentence.

In the early 1990, IBM developed a commercial Machine Translation system that used automatic bilingual text corpora and thus the idea of statistical approach was discarded. Statistical approach requires complex knowledge of mathematical equations and probabilities which was not very familiar to computational linguistic researchers.

Statistical approach underwent a lot of criticism as it was considered that notion of ‘probability of a sentence’ is a useless one. But one thing that researchers overlooked was that statistical approach dealt with the problem of taking decision and while developing an automatic speech recognition and text translation system taking decision was the major issue. So in the coming years the statistical approach became the widely accepted method for Machine Translation. So widely accepted that the leaders of technical industry like Google and IBM implemented this approach in there translators. Thus, then time came when traditional approached for machine translation like rule based became obsolete. Statistical approach brought an evolution to the computational linguistic industry as the error rate dropped to 20-25% compared to earlier approaches where the error rate was 52-60%. This was achieved only because statistical approach removed the black box concept and demanded heavy prior knowledge. This concept considerably reduced dependency on data thus; statistical approach achieved what was lacking in other approaches.

III. PROPOSED SYSTEM

Speech Recognition Process

In the speech recognition process, to get a meaningful text from the speech acoustic and language models act as the key players. The acoustic model determines the relationship between the audio from the microphone and the phonemes or other linguistic units that make up the speech. While Language model is responsible for modelling the word sequences in the language i.e. develop a meaningful text from the speech. As, we are reviewing a statistical system so we added both the models to represent the statistical properties of the speech.

In modern speech recognition systems the audio signal is distributed in frames. So the acoustic model does the sound analysis to model a relationship between the audio signals and the phonetic units by matching their patterns. Once this procedure is over the language model begins to work and develops the text by modelling the sequence in which the words of a particular language would be set so that the final text is meaningful. This process is explained vividly in Figure 1.

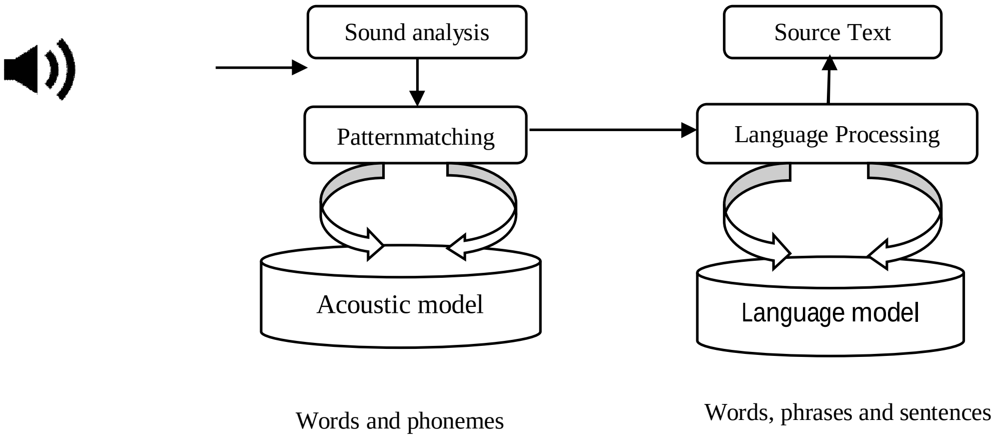

Figure. 1, Speech Recognition

Machine Translation Process

The main purpose of machine translation is to convert text in source language into target language. Let’s consider we have source string as $S_1^J = S_1 \ldots.. S_j \ldots.. S_J$ which is to be translated into a target string $T_1^I = T_1 \ldots.. T_j \ldots.. T_J$. Among all possible target strings, we will choose the string with the highest probability which is given by Bayes decision rule [6].

DOI: 10.15680/IJIRCCE.2016. 0405182

$$\hat{T}_1^I = \arg\max_{T_1^I} \{\Pr(T_1^I \mid S_1^J)\}$$

$$= \arg\max_{T_1^I} \{ \Pr(T_1^I) . \Pr(S_1^J \mid T_1^I)\}$$

Here, Pr (T_1^I) is the language model of the target language and Pr $(S_1^J \mid T_1^I)$ is the string translation model. The arg max operation denotes the search problem, i.e. the generation of the output sentence in the target language. The overall architecture of this translation approach is given in the following figure 2.

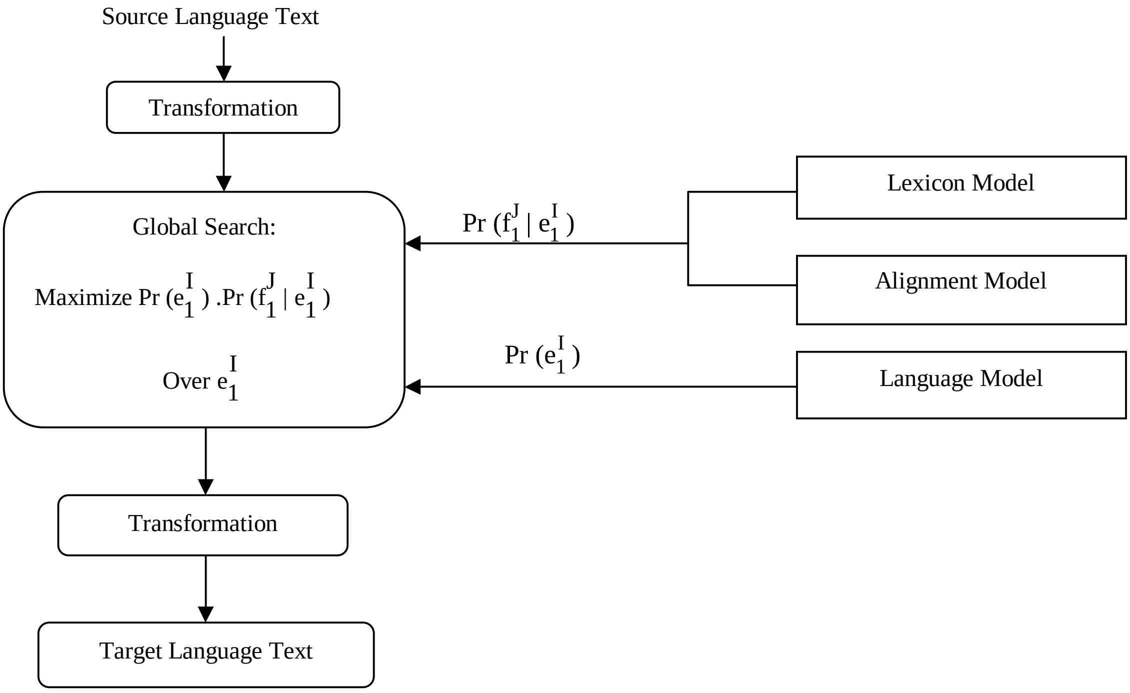

Figure 2, Machine Translation

Alignment models are used in statistical machine translation to determine translational correspondences between the words and phrases in a sentence in one language with the words and phrases in a sentence with the same meaning in a different language. They form an important part of the translation process, as they are used to produce word-aligned parallel text which is used to initialise machine translation systems. Improving the quality of alignment leads to systems which model translation more accurately and an improved quality of output.

Lexicon Model is a model which is extracted from word-aligned training data and—given the word alignment matrix—relies on pure relative frequencies. Lexical scoring on phrase level is the standard technique for phrase table smoothing in statistical machine translation. As most of the longer phrases appear only sparsely in the training data, their translation probabilities are overestimated when using relative frequencies to obtain conditional probabilities. One way to counteract overestimation of phrase pairs for which little evidence in the training data exists is to score phrases with word-based models and to interpolate these lexical probabilities with the phrase translation probabilities. Interpolation of the models is usually done loglinearly as part of the combination of feature functions of the translation system. In this way the interpolation parameter can be tuned directly against the metric of translation quality, on a held-out development set.In addition to phrase table smoothing, lexicon models are often applied on sentence level to re-rank the n-best candidate translations of the decoder. In re-ranking, the complete target sentence is available and the model can account for global sentence-level context to judge the selection of target words which was determined by the decoder. Both source-to-target and target-to-source models may be used.

Text-To-Speech (TTS)/Speech Synthesis

In this paper a novel approach is discussed to combine a rule based format synthesis approach and a corpus-driven approach. The new approach takes advantage of the fact that a unit library can better model detailed gestures then the current general rules. The rule based model keeps the flexibility to make modifications and the possibility to include both linguistic and extra linguistic knowledge sources. Figure 3 shows the approach from a technical point of view. A database used for creating a unit library. Each unit is described by a selection of extracted synthesis parameters with the linguistic information about the unit's original context and linguistic features. The parameters can be extracted automatically. In our traditional text-to-speech system the synthesiser is controlled by rule-generated parameters from the text-to-parameter module. These parameters are represented by time and values pairs including labels and prosodic features such as duration and intonation. In this approach some of the rule generated parameter values are replaced by values from the unit library. The process is controlled by the unit selection module that takes into account not only parameter information but also linguistic features supplied by the text-to-parameter module. The parameters are normalized and concatenated before being sent to the synthesizer.

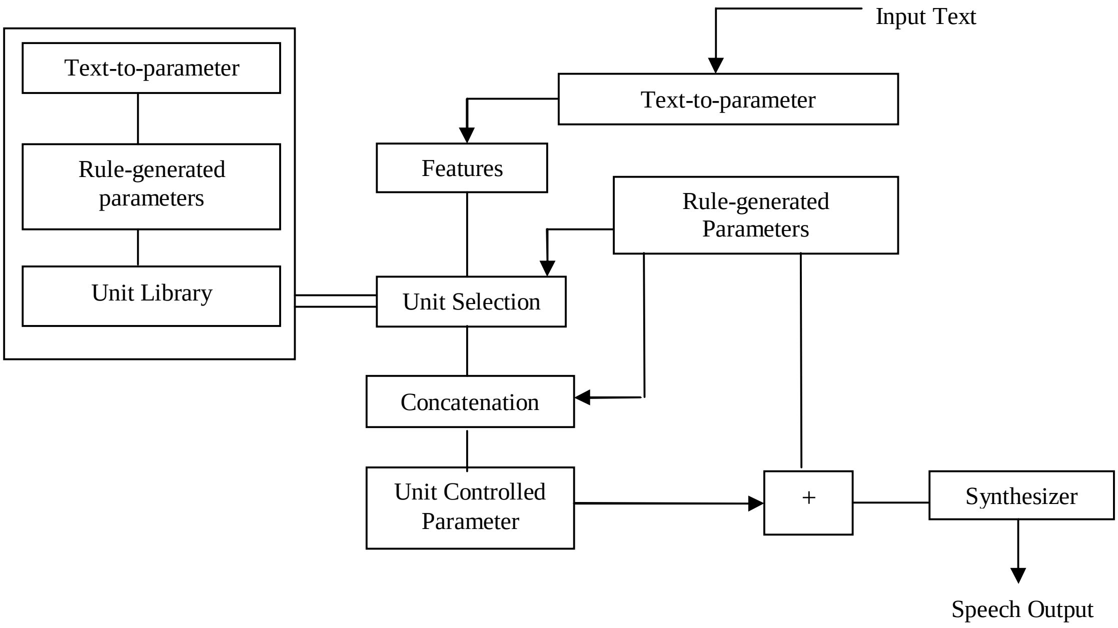

Figure 3, Speech synthesis system

For synthesis, the library requires three further three parts to make a complete synthesizer. Language model: providing phoneset, tokenization rules, text analysis, and prosodic structures etc. lexicon: a pronunciation model including a lexicon and letter to sound rules for out of vocabulary words. Voice- A voice depends upon the primitives provided by the language model.

IV. Proposed Algorithm

When we are dealing with linguistic and speech theories for computational purpose then we need to understand that both speech and linguistic units are not stationary units, rather they are quasi-stationary. So we cannot implement such models using static algorithms or algorithms based on fix parameters. Hidden Markov Model(HMM) is a the most powerful model that is best suited to implement all the above mentioned modules as they are all based on statistical approach. HMM is a tool in modern computational scenarios used to model wide range of time series data. It is the finest approach developed to deal with statistical variation of speech. Statistics of speech is assumed to differ from sample to sample over a short period of time. So to understand the HMM first we need to understand Markov Process which can be understood from figure 4.

Markov process is based on Markov property that can be used to model random system that changes its state based on a rule that is dependent only on the current rule. Thus, it is a stochastic model.

Figure 4 is description of a simple model for a stock market index. The model has three states, Bull, Bear and Even, and three index observations up, down, unchanged. The model is a finite state automaton, with probabilistic transitions between states. Given a sequence of observations, example: up-down-down we can easily verify that the state sequence that produced those observations was: Bull-Bear-Bear, and the probability of the sequence is simply the product of the transitions, in this case 0.2 × 0.3 × 0.3.

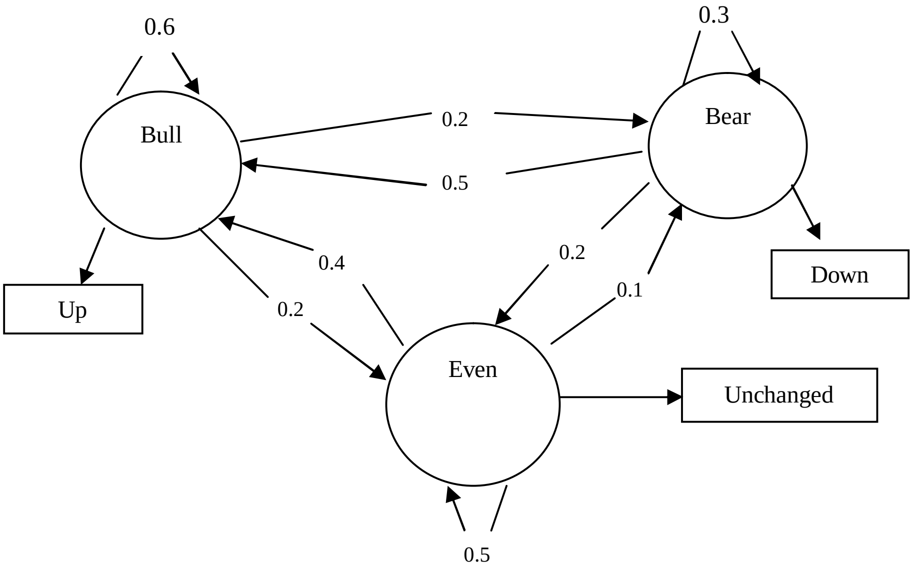

Figure 4, Markov Process

HMMs are used in speech related operations because a speech signal can be viewed as a piecewise stationary signal or a short-time stationary signal. In a short time-scales (e.g., 10 milliseconds), speech can be approximated as a stationary process. Speech can be thought of as a Markov model for much stochastic purposes. Another reason why HMMs are popular is because they can be trained automatically and are simple and computationally feasible to use. In speech recognition, the hidden Markov model would output a sequence of n-dimensional real-valued vectors (with n being a small integer, such as 10), outputting one of these every 10 milliseconds. The vectors would consist of cepstralcoefficients, which are obtained by taking a Fourier transform of a short time window of speech and decorrelating the spectrum using a cosine transform, then taking the first (most significant) coefficients. The hidden Markov model will tend to have in each state a statistical distribution which is explicated in Fig 1; shows the general speech recognition system. Each word, or (for more general speech recognition systems), mixture of diagonal covariance Gaussians, which will give likelihood for each observed vector. Each phoneme will have a different output distribution; a hidden Markov model for a sequence of words or phonemes is made by concatenating the individual trained hidden Markov models for the separate words and phonemes. So the combination of these models stores the knowledge of acoustic model, linguistic model as well as alignment model. Thus, a precise and a flexible model turn out to be the result.

V. Results

Table 1 depicts the evaluation of speech recognition under different environments for a stable result. The accuracy of the speech recognition module could only be justified after considering different environmental factors like noise.

An evaluation of the system accuracy for reference translations is shown in Table 3. The accuracies of the translation outputs ranked A (perfect), B (good), C (fair), or D (nonsense) by professional translators are shown in Table 3. The results match the BLEU scores shown in Table 2.

Characteristics	**No. of Speakers**			**No. of Utterances**			**Vocabulary Size (x10^3)**			**Perplexity**		
	English	Hindi	Bengali	English	Hindi	Bengali	English	Hindi	Bengali	English	Hindi	Bengali
Read Speech	8	8	8	320	320	320	47	41	38	17.8	22.4	31.1
Dialog speech in clean environment	5	4	2	314	313	314				21.6	27.3	43.7
Dialog Speech in Noisy Environment	6	4	5	104	103	101				24.3	20.9	40.5

Table 1. Evaluation of Speech Recognition

Language Pair	**BLEU**
English-to-Hindi	0.7012
Hindi-to-English	0.7489
English-to-Bengali	0.6531
Bengali-to-English	0.7200
Hindi-to-Bengali	0.5512
Bengali-to-Hindi	0.6579

Table 2. Objective Evaluation of Machine Translation Module

Language Pair	**Translation Accuracy (%)**			
	A	A+B	A+B+C	D
English-to-Hindi	77.3	87.3	94.2	7.6
Hindi-to-English	75.2	86.1	92.6	6.4
English-to-Bengali	67.7	77.8	89.6	12.5
Bengali-to-English	65.4	81.2	82.4	11.4
Hindi-to-Bengali	53.1	67.7	78.1	21.3
Bengali-to-Hindi	67.2	78.1	87.5	18.4

Table 3. Evaluation of the Translation Accuracy

VI. Conclusion and Future Work

This paper explained a novel technique for speech-to-speech translation using statistical approach. The hybrid of example base approach and statistical based approach gives an additional edge to the system that will help in reduction of error rate radically. We witnessed how the implementation of statistical approach in related systems reduced the error rate from 50-60% to 20-25%. Along with these approaches, the algorithm suggested in this paper called Hidden Markov Model (HMM) is one of the finest and most powerful modelsthat can be used to implement these approaches. This system will bring an evolution to the industry of computational linguistic. Globalization will reach its heights when language will no longer be a barrier. Although, a lot of features have to be added before we reach near perfection. Firstly, a better dictionary for the corpus and varied set of examples are needed, adding emotion to the speech synthesis is surely another feature that should be added as it will get easier to understand the context of the speaker.Not so soon but surely someday this system shall speak for us.

VII. Acknowledgment

I would like to thank Mousumi Ganguly, Sudip Ganguly and Akshay P. Deshmukh for their extensive support.

REFERENCES

1. Harold Somers, ‘Example-based machine translation’, Machine Translation, Vol.14, Issue 2. pp. 113-157, 1999
2. Ralf D. Brown, ‘Example-Based Machine Translation in the Pangloss System’, Association for computational linguistics, Vol. 1, pp. 169-174, 1996
3. Lambros CRANIAS, Harris PAPAGEORGIOU, Stelios PIPERIDIS, A Matching Technique in Example-Based Machine Translation, Association for computational linguistics, Vol. 1, pp. 100-104,1994
4. Franz Josef Och and Herman Ney, Statistical Machine Translation, In Final report, JHU summer workshop, Vol. 30, 1999
5. Parwinder Pal Singh and Er. Bhupinder Singh, ‘Speech Recognition as Emerging Revolutionary Technology’IJARCSSE, Vol. 2, Issue 10, 2012
6. F.Brown, S.A.DellaPietra, V.J. Della Pietra, and R.L Mercer. The mathematics of statistical machine translation: Parameter estimation, Computational Linguistics, Vol. 19, Issue 2, Pp. 263-311, 1993
7. James Brunning, Alignment Models and algorithms for statistical machine translation, Cambridge university Engineering department and Jesus College, Thesis submitted for Phd., August 2010
8. Phil Blunsom, ‘Hidden Markov Model’, Utah State University, Lecture Notes, Pp. 18-19, 2004
9. Zen Heiga , Takashi Nose , Junichi Yamagishi , Shinji Sako , Takashi Masuko , Alan W. Black , Keiichi Tokuda,The HMM-based speech synthesis system version 2.0, Proc. 6th ISCA Workshop on Speech Synthesis, SSW6, Pp. 294-299, 2007
10. Speech Interpolation for HMM-Based Speech Synthesis System, Takayoshi Yoshimura, Keiichi Tokuda ,Takashi Masuko, TakaoKobyashi, Tadashi Kitamura, Acoustical Science and Technology, Vol. 21(2000), Issue 4, Pp 199-206, 2001

BIOGRAPHY

I am an engineering graduate from Mumbai University in the field of Information Technology. Currently, preparing for my post-graduation and doing research on varied fields of Information Technology. I have also published another paper titled as ‘**A New Technique to Optimize User’s Browsing Session using Data Mining’**. I believe education and innovation are the absolute necessities for the growth of this world so I am trying my best to pursue my belief.