Journal of Engineering Science and Technology
Vol. 10, No. 2 (2015) 196 - 211

INTEGRATING MACHINE TRANSLATION AND SPEECH SYNTHESIS COMPONENT FOR ENGLISH TO DRAVIDIAN LANGUAGE SPEECH TO SPEECH TRANSLATION SYSTEM

J. SANGEETHA*, S. JOTHILAKSHMI

Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Chidambaram, India
*Corresponding Author: sangita.sudhakar@gmail.com

Abstract

This paper provides an interface between the machine translation and speech synthesis system for converting English speech to Tamil text in English to Tamil speech to speech translation system. The speech translation system consists of three modules: automatic speech recognition, machine translation and text to speech synthesis. Many procedures for incorporation of speech recognition and machine translation have been projected. Still speech synthesis system has not yet been measured. In this paper, we focus on integration of machine translation and speech synthesis, and report a subjective evaluation to investigate the impact of speech synthesis, machine translation and the integration of machine translation and speech synthesis components. Here we implement a hybrid machine translation (combination of rule based and statistical machine translation) and concatenative syllable based speech synthesis technique. In order to retain the naturalness and intelligibility of synthesized speech Auto Associative Neural Network (AANN) prosody prediction is used in this work. The results of this system investigation demonstrate that the naturalness and intelligibility of the synthesized speech are strongly influenced by the fluency and correctness of the translated text.

Keywords: Globalisation, Concatenative speech synthesis, Hybrid machine translation (HMT), Speech translation (ST), Subjective evaluation.

1. Introduction

The goal of the speech translation system research is to make straightforward real-time, interpersonal communication via usual spoken language for people who do not share a neutral language. Speech Translation (ST) is the process [1] by

which spoken expressions are rapidly translated and spoken clearly in a second language. This is in contrast from phrase translation method, where the system merely translates a predetermined and finite set of sentences that have been manually entered into the system. Speech to speech translation technology supports speakers of various languages to interconnect. Thus it provides fabulous value for humankind in terms of science, cross-cultural interchange and global business. Nowadays, speech translation systems are used all through the domain. Examples include medical facilities, police, schools, retail stores, hotels, and factories. These systems are applicable everywhere that spoken language is being used to communicate.

Currently, speech translation technology is available as a product that instantly interprets free form multi-lingual conversations. These systems instantly convert uninterrupted speech. Challenges in achieving this task include overcoming speaker dependent changes in fashion of speaking or pronunciation are issues that have to be dealt with so as to give a high quality translation for every user. Moreover, automatic speech recognition systems have to be prepared to tackle external factors such as acoustic noise or speech by other speaker in real-world use of speech translation systems. For the motivation that the client does not know the target language when speech translation is used, a method need be delivered to the user to check whether the translation is correct, by such means as translating it again back into the user's language [2].

In order to break the goal of wiping out the language barrier worldwide, multiple languages ought to be supported. This requires speech corpora for each of the calculated 6,000 languages said to exist on our globe today. The corpora collection is extremely expensive and gathering data from the Web would be a substitute to conventional methods. Secondary use of news or other media bring out in multiple languages would be an effective way to progress performance of the system translation system. So far, up to date copyright law does not yield secondary uses such as these types of corpora into account and thus it will be necessary to review it so that it is additional flexible.

A speech to speech translation system [3] comprises three components: speech recognition, machine translation and speech synthesis. In case of a simple S2ST system, only the single-best output of one component is used as input to the next component. As a result, the errors of the previous module strongly affect the performance of the next module. Due to the slips in automatic speech recognition module, the machine translation component cannot accomplish the equivalent level of translation performance as accomplished for correct text input. In order to overcome this issue, numerous techniques for incorporation of speech recognition and machine translation have been proposed. When coupling automatic speech recognition and machine translation components, the recognition module scores and the translation module scores can be pooled to increase translation performance.

A theoretical basis for the score combination was given in [4, 5]. In these, the impact of speech recognition errors in machine translation is alleviated by using word lattice or *N*-best list output from the speech recognition component as input to the machine translation element. Therefore, these methodologies can progress the performance of S2ST considerably. Though, the speech synthesis element is not typically considered. The output speech signal for translating sentences is produced by the speech synthesis component. If the quality of synthesized speech

is corrupt, end users cannot recognise what the system said. Thus the quality of synthesized speech is apparently essential for S2ST and any integration method projected to improve the end- to-end performance of the system should take justification of the speech synthesis component.

The EMIME project [6] is emerging personalised S2ST, such that the user's speech input in one language is used to yield output speech in another language. Speech characteristics of the produced output speech are adjusted to the input speech characteristics using cross-lingual speaker adaptation methods (EMIME project). While personalisation is a significant area of research, this paper focuses on the impact of the machine translation and speech synthesis components of an S2ST system. In order to recognise the degree to which each component affects performance, we examined integration methods. We first directed a subjective evaluation which comprises three sections: speech synthesis, machine translation, and integration of MT and TTS. Various sentences were evaluated by using *N*-best translated sentences output from the machine translation component. The individual impressions of the machine translation and the speech synthesis modules are investigated from the results of this subjective evaluation.

2. Related Work

The quality of synthesized speech is one of the most eminent features in the field of spoken dialog systems because users cannot recognise what the system intended to say if the quality of synthesized speech is low. As a result, incorporation of natural language generation (machine translation) and speech synthesis has been proposed [7-9].

A method was proposed [7] for integration of automatic machine translation and unit selection based text to speech synthesis system which allows the selection of wording and prosody to be jointly determined by the machine translation and speech synthesis components. A template-based language generation module passes a word network expressing the same content of the speech synthesis component, rather than a single word string. To attain the unit selection search on this word network input effectively, weighted finite-state transducers (WFSTs) are employed. The weights of the WFST are determined by prosodic prediction costs, join costs, and so on. In an experiment, this system accomplished higher quality speech output. This technique cannot be used with most prevailing speech synthesis systems, because they will not admit word networks as input. A substitute to the word network approach is to re-rank sentences from the *N*-best output of the natural language generation component [8]. *N*-best output can be used in combination with any text to speech synthesis system even if the natural language production section must be able to construct *N*-best sentences. In this method, a re-ranking model select the sentences that are anticipated to sound more natural when synthesized with the unit selection based speech synthesis module.

The re-ranking model is skilled from the subjective scores of the produced synthesize speech quality assigned in a preliminary estimation and features from the machine translation and speech synthesis components such as word N-gram model scores, join cost, and prosodic prediction costs. Tentative results revealed higher quality speech output. Similarly, a re-ranking model for N-best output was also

proposed in [8]. In contrast this model used a much smaller data set for training and a larger set of features, however reached the same performance as reported in [9].These incorporation methods are in natural language generation and text speech synthesis for verbal dialog systems. On the contrary to these methods, our focal point is on the integration of automatic machine translation and text to speech synthesis system for speech to speech translation system. To this end, we first conducted a subjective evaluation independently for proposed hybrid machine translation and syllable based TTS system, and then analysed the impact of integrated machine translation and speech synthesis system for S2ST system.

This paper describes the recent progress in the research on integrating machine translation and speech synthesis system for developing English text to the Tamil speech synthesis system, including the development of English to Tamil hybrid machine translation system (combination of rule based and statistical machine translation system) and syllable based concatenative speech synthesis system. Firstly we describe the hybrid machine translation system in Section 3. Section 4 discusses the development of syllable based concatenative text to speech synthesis system. The integration of these component technologies for English text Tamil speech synthesis system is described in Section 5. Section 6 discusses the objective and subjective performance measures of the developed components. Finally, we draw our conclusion in Section 7.

3. Hybrid Machine Translation (HMT) System

A number of machine translation systems have been proposed in the literature. But, conventional rule-based machine translation system is costly in terms of formulating rules. It introduces inconsistencies, and it is inflexible to be robust. Statistical MT is an approach that automatically attains knowledge from a vast amount of training data. This approach is characterised by the use of machine learning techniques. But, still there is scope for better performance of the system.

In this paper, a Hybrid Machine Translation (HMT) approach is proposed which is the combination of rule based and statistical technique for translating text from English to Tamil. The rule based machine translation technique, involves the formation of rules which helps in re-ordering of the syntactic structures of the source language sentence along with its dependency information which bring close to the structure of the target sentence. The parser categorises the syntactical essentials in English sentences and proposes its Indian languages translation taking into consideration various grammatical forms of those Indian languages. Context Free Grammars (CFG) is used in generation of the language structures, and then the errors in the translated sentences are corrected by applying a statistical technique. Simplifying and segmenting an input language text becomes mandatory in order to improve the machine translation quality.

Rule based machine translation for a language is done by developing the rule structures for every possible sentence in the language. The grammar and dictionary creation are the main tasks in the rule based machine translation. The grammar rules and dictionaries once created can be used and never need changing. In order to improve the machine translation quality, splitting and simplifying an input becomes mandatory. The overview of the proposed machine translation system is shown in Fig. 1.

3.1. Segmentation

The given input text is simplified first. We are using a rule based technique [10] to simplify the paragraph and complex sentences. It is based on coordinating conjunctions (for, and, not, but, or, yet and so), subordinating conjunctions (after, although, because, before, if, since, that, though, unless, where, wherever, when, whenever, whereas, while, why) and connectives like relative pronouns (who, which, whose, whom).

3.2. POS (Parts Of Speech) tagging

The tagging of the source sentence is done using the Parts Of Speech tagger. POS is more important in case of words because it gives information about the word's pronunciation and the words morphological affixes. Tag sets are mostly language independent. It assigns the same tag to all words.

The English sentences are fed into the parts of speech tagger. The output of the vocabulary and their tagged information are stored in a separate place which is utilised for reordering according to the Tamil language organisation. Here we are using Standford POS tagger for tagging the information.

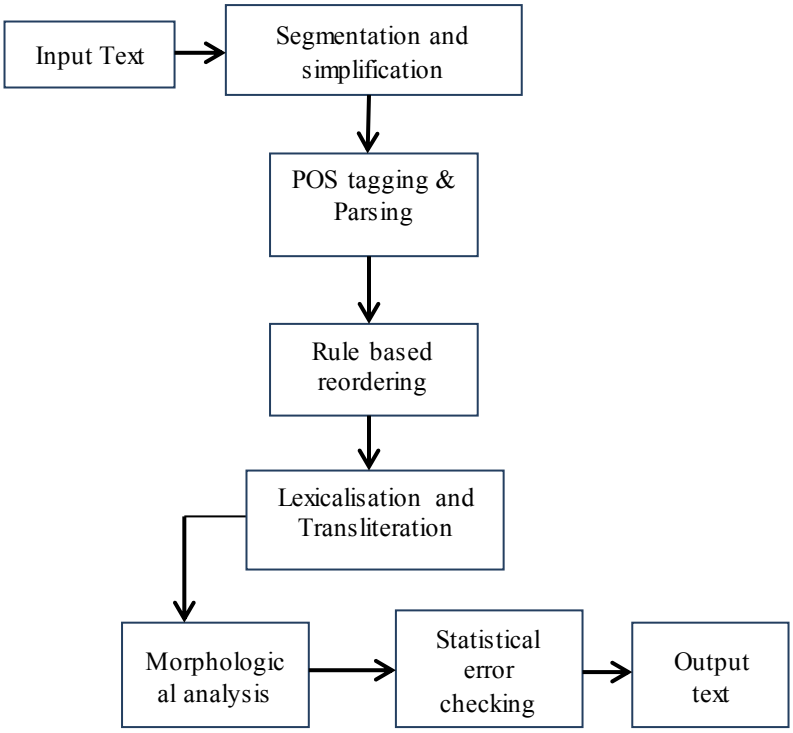

Fig. 1. Overview of the proposed hybrid machine translation system.

3.3. Rule based reordering

Reordering is the ultimate need for English to any Indian language machine translations. The major structural difference between English and Indian languages is that English follows structure as Subject-Verb-Object (SVO), whereas Indian languages follow default sentence structure as Subject-Object-Verb (SOV).

With the differences of word order between English and Indian languages, handling absolutely the reordering problem is necessary Consider the following sentence "He went to shop".

Source sentence: He went to the park

Target sentence:Avan poongavirkku sendraan(Tamil)

In this example, SVO in source sentence is transformed as SOV in target sentence

He : Subject : Avan

Park : Object : poongavirkku

Went to : Verb : sendraan

This reordering is done according to the pre-specified rules. These rules are handcrafted by both the linguists and computer programmers. These rules are present inside the database. In the above sentence, the general structures for the source side and target side are shown in Fig. 2.

Sentence: He went to the park.	
Source Side :	S -> NP VP VP -> VBD S
Target side:	S -> NP VP VP -> S VBD

Fig. 2. Reordering rules for the sentence 'He went to shop'.

3.4. Lexicalisation

For the purpose of mapping words from one language to another, we require a lexical dictionary. The lexical dictionary contains the word in the source language and its corresponding root word in target language. Our dictionary contains three different tables. In Tamil, gender information plays an important role in the formation of a word. Sentence formation is dependent upon the gender of the subject. Hence it is important to add information about gender in the dictionary. Certain words which are considered to be singular in English are considered plural in Tamil and vice versa.

3.5. Transliteration

Transliteration can be described as "systematic transliteration is a mapping from one structure of inscription into another, word by word, or supremely letter by letter". Transliteration is implemented using the rule based procedure.

Recognizing the gender information for the transliterated proper nouns is an unattainable task. Hence, for any proper noun that is transliterated, the common gender ending in Tamil language for both male and female gender "kiraar" is applied. Let us consider an example sentence **"**Sairam is driving a car". For this the translated output sentence is "Sairam kaar ooti kondirukiraar"

3.6. Morphological analyser

Once the sentences are translated into the target language the morphological analysing has to be done. Because, in Tamil language a single word for example "selkirran" shows the tense, gender, and the action performed which appears like a complete sentence. So the gender ending with the proper tenses has to be added with the verbs. From the input sentence, we have to extract the following information such as tense information, subject and verb in the sentence. The subject and the verb in the sentence can be easily identified using the dependency relation given by the Parser.

Consider the example given below: The output of the sentence "He is going to shop" after the reordering process will be "**He shop to going is**". At this point the "**going**" is the verb in the continuous tense. Here the root word "going" has to be originating which is "**go**". After getting the root word it will be simple to locate the meaning in the dictionary. With the help of the tagged information, the gender ending and tense marker processes are completed.

4. Syllable Based Concatenative Text to Speech Synthesis using Auto Associative Neural Network (AANN) Prosody Prediction

Unrestricted Tamil Text To Speech System (TTS) is capable of synthesizing different domain speech with improved quality. In this work a corpus-driven Tamil text-to-speech system based on the concatenative synthesis approach is developed. Concatenative speech synthesis involves the concatenation of the basic units to synthesize an intelligent, natural sounding speech. A corpus-based method (unit selection) uses a large inventory to select the units and concatenate. The prosody prediction is done with the help of fiver layer auto associative neural network which helps us to improve the quality of speech synthesis. Here syllables are used as basic unit of speech synthesis database. The database consisting of the units along with their annotated information is called annotated speech corpus. A clustering technique is used in annotated speech corpus that gives the way to choose the suitable unit for concatenation, based on the lowest total joint cost. Discontinuities present at the unit boundaries are lowered by using the Mel-LPC smoothing [11] technique. The experimental results reveal to demonstrate the improved intelligibility and naturalness of the proposed method.

Concatenative synthesis approach [12] is very complex to build up because of restriction in computer memory. By means of the improvement in computer hardware and memory management, a huge amount of speech corpus can be stored and used to produce high quality speech waveforms for a given text. Thus, the synthesized speech waveform preserves the naturalness and intelligibility. Concatenative speech synthesis is depends on the concatenation of small speech fragments of recorded speech. In general, concatenative synthesis system yields

the supreme natural-sounding synthesized speech. Concatenative speech synthesis employs phones, diphones, syllables, words and sentences as fundamental units. Speech is synthesized based on choosing these elements from the database, called a speech corpus.

Indian languages are syllable oriented, where pronunciations are based on syllables. A syllable can be the finest unit for Indian languages intelligible speech synthesis system. The general form of Indian language syllable is C*VC*, where C is a consonant, V is vowel and C* indicates the occurrence of 0 or more consonants. There are 18 consonants and 13 vowels in Tamil languages. There are defined set of syllabification rules conversion [13] formed by researchers, to produce computationally reasonable syllables.

4.1. Speech corpus

Building a speech corpus for Tamil language is a more difficult task than for English speech corpus. For developing the corpus for TTS system, we have considered 500 Tamil language sentences. These 500 sentences are split into syllables based on linguistic rules and the syllables are recorded using two female and three male persons, whose age is in the range of 25–35 years. The reason for recording the speech syllables with multiple speakers is to select the preeminent speaker for recording the final speech to develop the unrestricted and full-fledged TTS. In this TTS corpus creation, best speaker means the speech produced by that speaker who has uniform characteristics with respect to speaking rate, intonation, energy profile and pronunciation. In concatenative TTS, the quality of speech corpus is very important, because the characteristics of synthesized speech are directly related to nature of speech corpus.

Along with the above mentioned characteristics, the speech of the selected speaker should give least distortion when speech segments are manipulated as per requirements. First, 500 Tamil sentences syllables are recorded by each of the five speakers in an echo free audio recording studio. The recorded speech signal is sampled at 48 kHz and stored in 16-bit PCM data format. After recording the speech from each speaker each of the speech file is down sampled to 16 kHz. The speech wave files are saved according to the requirement. The speech wave files corresponding to the Tamil words are named according to their corresponding Romanised names. Each syllable and word file contains text transcriptions and timing information in number of samples. The fundamental frequencies of the syllables are computed using autocorrelation method. The words collected comprise dictionary words, commonly used words, Tamil newspapers and story books, also different domains such as sports, news, literature and education for building unrestricted TTS [14].

4.2. Prosody modelling with auto associative neural networks

Syllable level prosodic parameters of speech signal depend upon positional, contextual or background and phonological features of the syllables [15]. In this paper, auto associative neural networks are in a job to represent the prosodic parameters of the syllables from the features. The prosodic parameters measured in this work are duration and sequence of pitch (F0) values of the syllables. This

technique is used in text to speech synthesis process. A five-layer auto associative neural network shown in Fig. 3 is used for modelling the duration/F0 patterns of syllables [16]. The structure of the AANN is show in Fig .3.

In this paper, we consider the mapping function is between the 25-dimensional input vector and the one-dimensional output. Several network structures are explored in this study. The final structure of the network is 25L 50N 5N 50N 25N, where L denotes a linear unit, and N denotes a non-linear unit. For modelling the durations of syllables, neural network models are urbanised with respect to the Tamil language. In the case of intonation modelling, a separate model is developed for each of the speakers in Tamil language. For each syllable, a 25-dimension input vector is formed, representing the positional, contextual and phonological features.

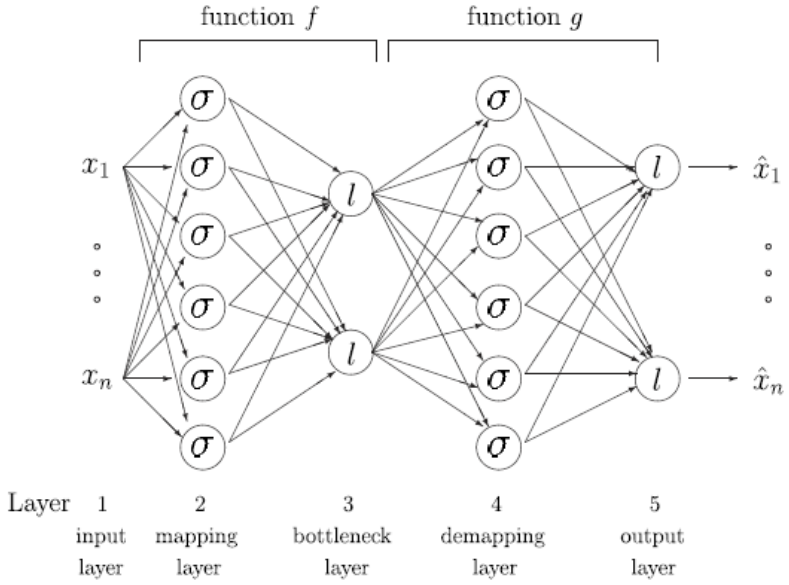

Fig. 3. A Five-layer auto associative neural network model for capturing the duration/F0 patterns of syllables.

To evaluate the prediction accuracy, the average prediction error (μ), the standard deviation (σ), and the correlation coefficient (γ X, Y) are computed using actual and predicted *F*0/duration values. The definitions of average prediction error (μ), standard deviation (σ) and the linear correlation coefficient (γ X, Y) are given below.

$$\mu = \frac{\sum_i | x_i - y_i |}{N} \tag{1}$$

$$\sigma = \sqrt{\frac{\sum_i d_i^2}{N}} \tag{2}$$

$$d_i = e_i - \mu \tag{3}$$

$$e_i = x_i - y_i \tag{4}$$

where x_i, y_i are the actual and predicted $F0$ values, respectively, and e_i is the error between the actual and predicted $F0$ values. The deviation in error is d_i, and N is the number of observed $F0$ values of the syllables. The correlation coefficient is given by:

$$\gamma_{x,y} = \frac{V_{x,y}}{\sigma_x, \sigma_y} \tag{5}$$

where

$$V_{x,y} = \frac{\Sigma_i |(x_i - \bar{x})| . |(y_i - \bar{y})|}{N} \tag{6}$$

The quantities σ_X, σ_Y are the standard deviations for the actual and predicted $F0$ values, respectively, and $V_{X,Y}$ is the correlation between the actual and predicted $F0$ values.

To estimate the prediction accuracy, the average prediction error (μ), the standard deviation (σ), and the correlation coefficient (γ X, Y) are computed using actual and predicted $F0$/duration values. These results are given in Tables 1 and 2.

4.3. Proposed text to speech synthesis system

The block diagram of the proposed unrestricted AANN based TTS for Tamil is shown in Fig. 4. In the front end, the first step is text analysis which is an automated process of removing punctuations such as double quotes, full stop, comma and all. The analysed text is syllabified based on the linguistics rules.

Table 1. Performance of the prosody models for predicting the *F*0 of the syllables N= 1276 syllables.

Predicted syllables within deviation				Objective measures		
5%	10%	15%	25%	μ (ms)	σ (ms)	γ
41	74	92	99	16.02	13.04	0.8

Table 2. Performance of the prosody models for predicting the durations of the syllables.

Predicted syllables within deviation			Objective measures		
10%	25%	50%	μ (*ms*)	σ *(ms)*	γ
36	78	98	27	24	.83

Then the positional, contextual and phonological features (25 dimensional) for each of the syllables present in the given text are derived. These features are given to the prosody prediction models (duration and intonation model) which will generate the appropriate duration and intonation information corresponding to the syllables. At the synthesis stage, first, the concatenation is performed based on the

pre-recorded syllables according to the sequence in the text. Using prosody modification methods the derived duration and intonation knowledge corresponding to the sequence of syllables is incorporated in the sequence of concatenated syllables. The main problem in concatenation process is that there will be glitches in the joint. These discontinuities present at the unit boundaries are lowered by using the Mel-LPC smoothing technique. This is one of the most dominant methods for encoding high-quality speech at a low bit rate and offers tremendously exact estimates of speech parameters [17, 18].

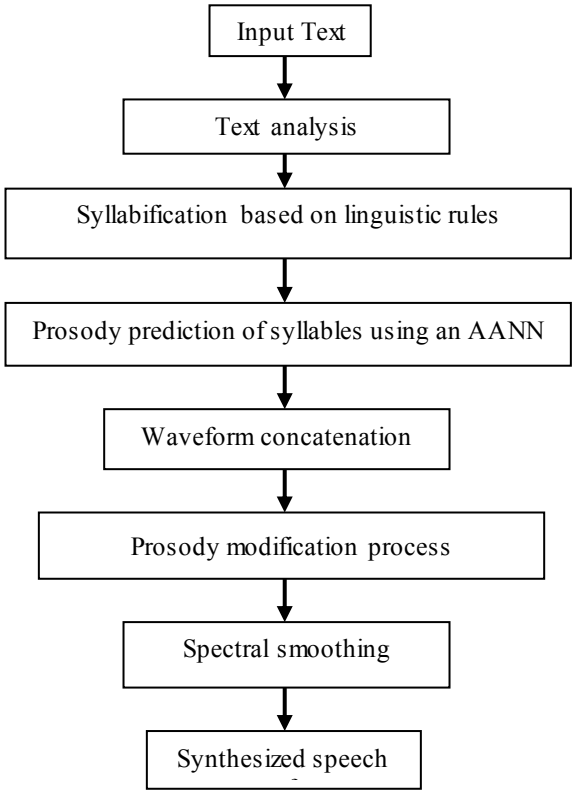

Fig. 4. Block diagram of the proposed Tamil text to speech synthesis system.

5. Integrating Machine Translation and Speech Synthesis System

The incorporated English to Tamil hybrid machine translation system and concatenative syllable based TTS are shown in Fig. 5. It is designed for practical use of English to Tamil speech to speech translation system for various domains. Fig. 5 shows the overall architecture of integrating HMT and TTS. It consists of machine translation module and speech synthesis module. This system will translate the English text to Tamil text and produces the synthesized Tamil speech with naturalness and intellectual for the given English text.

The project aims at upgrading the development of multilingual man-machine interfaces, particularly the multilingual speech translation systems in the Indian regional states. Thus, the final system is expected to include the Automatic

Speech Recognition (ASR) module in order to make the complete speech to speech translation system, and to also include the other Indian languages such as Hindi, Malayalam, Kannada, etc.

In this proposed integration of HMT and TTS, the correct sentences were used as the input of the machine translation component instead of speech recognition results. The machine translation module takes the input and performs the translation process. The MT provides the exact translation results because of the combination of RBMT and SMT. The translated Tamil input text is fed to TTS module in order to produce the good quality synthesized speech. In TTS, AANN based prosody prediction is used so that it provides the speech in Tamil with naturalness and intelligent. And also concatenated TTS added advantage to produce the virtuous quality speech waveforms.

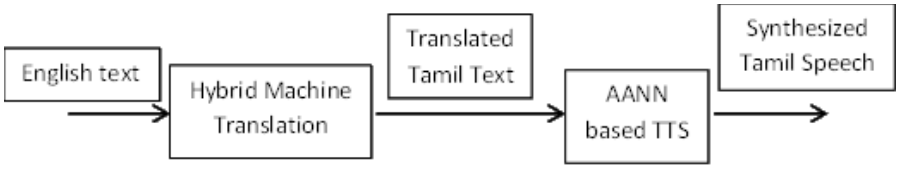

Fig. 5. Integrating HMT and TTS.

6. Result Evaluation

The evaluation comprised 3 sections; In section 1, hybrid machine translation was evaluated. In section 2 speech synthesis system was evaluated. In section 3, integrated HMT and TTS were evaluated based on the given input to MT and the corresponding synthesized speech output.

6.1. Evaluation procedure for HMT

The automatic evaluation for machine translations is BLEU (Bilingual Evaluation Under-Study). BLEU [19] metric gives high correlation against human judgment and it is one of the popular metrics in machine translation since it gives maximum correlation with the human translation. BLEU calculates the average score of the individual sentences to get the final score for the whole corpus. To compare the candidate translation and machine translation, BLEU uses modified form [20] of precision because machine translation system generates more words than they actually appear in the reference text.

Our ultimate goal is to provide better accuracy for the translation system. 200 sentences in each language are given to the hybrid machine translation out of which 140 sentences are given incorrect because of syntax and reordering errors. But the same 200 sentences were tested after segmentation and simplification. 115 sentences are translated correctly with simplification. The system provides good results because the input is simplified and segmented before enter in to the translation process, rule based reordering is performed and finally the errors are corrected statistically. Proposed hybrid machine translation system was evaluated using BLEU evaluation metric. Table 3 summarises the comparison of results obtained from the proposed work with the baseline system.

'Baseline' stands for simple Hybrid machine translation system results; 'Baseline+ Simplification' stands for the results after simplification and segmentation. 'Baseline+ Simplification + Syntax stands for the results after simplification and reordering. 'Baseline+ Simplification + Syntax + Morphology' stands for morphological processing followed by re-ordering and segmentation. In all the cases training was performed with the same corpora and the same set of sentences was used for testing. The blue scores obtained for the basic hybrid system were low as the training corpus size that used was very less and obviously depends on the testing data that we used. The performance evaluation shows that when we applied segmentation and simplification, reordering and adding morphological information the blue score increased by approximate 3.0 to 5.5 for each language. Thus, the proposed hybrid machine translation system exhibits the improved performance for the translation from English to Tamil language.

Table 3. Evaluation statistics of proposed HMT.

Technique	BLEU score for Tamil
Baseline	15.5
Baseline + Simplification	18.2
Baseline + Simplification+Syntax	21.9
Baseline +Simplification+Syntax+Morphology	24.9

6.2. Subjective evaluation of TTS

Voice quality testing is performed using subjective test. In subjective tests, human spectators perceive sound and grade the quality of processed voice files according to a certain scale. The most widespread scale is called Mean Opinion Score (MOS) and is self-possessed of 5 scores of subjective quality, 1-Bad, 2-Poor, 3-Fair, 4-Good, 5-Excellent. The MOS score of a certain TTS system is the average of all the ranks voted by different listeners of the different voice file used in the experiment. The tests were conducted in a laboratory environment with 50 students in the age group of 20-28 years by playing the synthesized Tamil speech signals through headphones. In this case, the subjects should possess the adequate speech knowledge for accurate assessment of the speech signals and were examined to evaluate the articulacy and spontaneity of the synthesized speech. They have to assess the quality on a 5-point scale for each of the sentences. The mean opinion scores for assessing the intelligibility and naturalness of the synthesized Tamil speech are given in Table 4.

The MOS scores show that the intelligibility of the synthesized Tamil speech is honestly acceptable, whereas the naturalness appears to be little degree of degradation. Naturalness is mostly attributed to distinct perception. It can be enhanced to some degree by integrating the stress and co articulation information along with duration and pitch. The accuracy of the prediction of prosody models can be also analysed by conducting the listening tests for judging the intelligibility and naturalness on the synthesized speech without incorporating the prosody. In this case, speech samples are the derivative of concatenating the neutral syllables without integrating the prosody.

The MOS of the excellence of the synthesized speech without incorporating the prosody have been observed to be low compared to the speech synthesized by

combining the prosody. The consequence of the differences in the pairs of the MOS for intelligibility and naturalness is verified using hypothesis testing and the level of confidence is high (>99.5%) for both cases.

Table 4. Mean opinion score for the quality of synthesized speech in Tamil language.

Mean opinion score				Level of confidence (%)	
TTS with prosody		TTS without prosody			
Intelligence	Naturalness	Intelligence	Naturalness	Intelligence	Naturalness
4.5	4.2	3.9	3.1	> 99.5	> 99.5

6.3. Evaluation of integrated HMT and TTS

Evaluators listened to synthesized speech and the corresponding typed text in HMT. Word Error Rate (WER) is measured here. After this, evaluators assigned scores for "Adequacy" and "Fluency" of the typed-in sentence. Here, "Adequacy" indicates how much of the information from the reference translation sentence was expressed in the sentence and "Fluency" indicates how fluent the sentence was [21]. These definitions were provided to the evaluators. Evaluators assigned scores to 50 test sentences in each section. 100 people participated in the evaluation. "Adequacy" and "Fluency" measures can be evaluated by monolingual target language listeners. These measures are widely used in machine translation evaluations, e.g., conducted by NIST and IWSLT.

We analysed the impact of the translated sentences on the naturalness and intelligibility of synthesized speech. Table 5 shows the correlation coefficients between Text to speech synthesis and automatic machine translation scores, and the correlation coefficients between word error rate and machine translation scores. MT-Fluency score has a rigid correlation with both TTS score and WER than MT-Adequacy score.

Next we focused on the relationship between the fluency of the machine translation output and the corresponding synthesized speech. Speech synthesis text often produced incorrect full-context due to the errors in syntactic analysis of disfluent and ungrammatical machine translated sentences.

Table 5. Correlation coefficients between TTS or WER and MT.

	MT-Adequacy	MT-Fluency
TTS	0.14	0.27
WER	-0.18	-0.24

In addition, the psychological effect called "Llewelyn reaction" [22] seems to modify the results of the synthesized speech. The "Llewelyn reaction" is that evaluators notice lesser speech quality when the sentences are less fluent or the content of the sentences is less natural, even though the authentic quality of synthesized speech is unchanged. Consequently, we conclude that the speech synthesis component will lean towards to produce more natural speech as the translated sentences become more fluent.

The impact of WER in MT-Fluency is assumed to affect the synthesized speech because the evaluators can predict the next word when unusual words or phrases are not in the translated sentence. More over the naturalness of synthesized speech is enhanced when the sentences are more fluent. Naturalness and intelligibility of synthesized speech waveforms are inclined by the fluency of the machine translated sentences than by the content of sentences. Hence, the intelligibility and naturalness of the synthesized speech is upgraded as the translated sentences become more fluent, even though all sentences are synthesized by the same system.

7. Conclusion

This paper has provided an integration of hybrid machine translation and syllable based speech synthesis system for English to Tamil speech-to-speech translation system and also addressed the issues that arise in the integration of MT and TTS components. Experimental results show that significant translation accuracy enhancements can be achieved by paying special attention to sentence segmentation, reordering and morphological analysis. Similarly the hybrid machine translation approach considerably aggregates the translation accuracy. We have revealed that the naturalness and intelligibility of the synthesized speech waveforms are severely exaggerated by the fluency of the machine translated texts. The intelligibility of synthesized speech is improved as the translated sentence becomes more fluent. In addition, prosody predictions of the syllables using AANN in TTS were used to enhance intelligibility and naturalness. Our future work will include investigations into integrating Automatic Speech Recognition (ASR) into this MT & TTS interfacing and also joint optimisation of ASR, TTS and MT components using translation of speech lattice output, word N-gram and phoneme N-gram scores.

References

1. He, X.; and Deng, L. (2011). Speech recognition, machine translation and speech translation - a unified discriminative learning paradigm. *IEEE Signal Processing Magazine*, 28(5), 126-133.
2. Satoshi. (2009). Overcoming the language barrier with speech translation technology. *Nakamura in Science & Technology Trends - Quarterly Review*, 31, 35-48.
3. Hashimoto, K.; Yamagishi, J.; William, B.; Simon, K.; and Keiichi, T. (2011). An analysis of machine translation and speech synthesis in speech-to-speech translation system. *International Conference on Acoustics, Speech and Signal Processing*, 5108-5111.
4. Vidal, E. (1997). Finite-state speech-to-speech translation. *International Conference on Acoustics, Speech and Signal Processing*, 111-114.
5. Ney, H. (1999), Speech translation: coupling of recognition and translation. *International Conference on Acoustics, Speech and Signal Processing*, 1149-1152.
6. Simon, K. (2015). Retrieved January 14, 2015 from http://www.emime.org/

7. Wu, Y.-J.; Nankaku, Y.; and Tokuda, K. (2009). State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis. *Proceeding of Interspeech Conference*, 528-531.
8. Bulyko, I.; and Ostendorf, M. (2002). Efficient integrated response generation from multiple target using weighted finite state transducers. *Computer Speech and Language*, 16(3-4), 533-550.
9. Nakatsu, C.; and White, M. (2006). Learning to say it well: Reranking realisations by predicted synthesis quality. *International Proceeding of the Annual Meeting of the Association for Computational Linguistics*, 1113-1120.
10. Boidin, C.; Rieser, V.; Plas, L.V.D.; Lemon, O.; and Chevelu, J. (2009). Predicting how it sounds: Re-ranking dialogue prompts based on TTS quality for adaptive spoken dialogue systems. *Proceeding of Interspeech Conference*, 2487-2490.
11. Varho, S.; and Alku, P. (1997). Linear predictive method using extrapolated samples for modelling of voiced speech. *In Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 13-16.
12. Poornima, C.; Dhanalakshmi, V.; Anand, K.M.; and Soman, K.P. (2011). Rule based sentence simplification for English to Tamil machine translation system. *International Journal of Computer Applications*, 25(8), 38-42.
13. Lokesh, S.; and Balakrishnan, G. (2012). Speech enhancement using Mel-LPC Cepstrum and vector quantisation for ASR. *European Journal of Scientific Research*, 73(2), 202-209.
14. Shirbahadurkar, S.; Bormane, D.; and Kazi, R. (2010). Subjective and spectrogram analysis of speech synthesizer for Marathi's using concatenative synthesis. *International Conference on Recent Trends in Information, Telecommunication and Computing*, 262-264.
15. Saraswathi, S.; and Geetha, T.V. (2010). Design of language models at various phases of Tamil speech recognition system. *International Journal of Engineering, Science and Technology*, 2(5), 244-257.
16. Robert, J.; Utama, Ann.; Syrdal, K.; and Alistair, C. (2006). Six approaches to limited domain concatenative speech synthesis. *Inter Speech*.
17. Yegnanarayana, B. (1999). *Artificial neural networks*. Prentice-Hall, New Delhi.
18. Sreenivarao, K. (2011). Role of neural network models for developing speech systems. *Indian Academy of Sciences Sadhana,* 36(5), 783-836.
19. White, J.S.; Connell, T.O.; and Mara, F.O. (1994). The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. *International Proceeding AMTA*, 193-205.
20. Yang, M.Y.; Zhu, J.G.; Li, J.F.; Wang, L.X.; Qi, H.L.; Li, S.; and Liu, D.X. (2008). Extending BLEU evaluation method with linguistic weight. *International Conference for Young Computer Scientists, ICYCS*, 1683-1688.
21. Varho, S.; and Alku, P. (1997). Linear predictive method using extrapolated samples for modelling of voiced speech. *In Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 13-16.
22. Yamada, S.; Kodama, S.; Matsuoka, T.; Araki, H.; Murakami, Y.; Takano, O.; and Sakamoto, Y. (2005). A report on the machine translation market in Japan. *In Proceeding MT Summit X*, 55-62.