IOSR Journal of VLSI and Signal Processing (IOSR-JVSP)
Volume 6, Issue 6, Ver. I (Nov. - Dec. 2016), PP 77-84
e-ISSN: 2319 – 4200, p-ISSN No. : 2319 – 4197
www.iosrjournals.org

A Review on Sub-word unit Modeling in Automatic Speech Recognition

Karpagavalli S[1], Chandra E[2]
[1](Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore, India)
[2](Department of Computer Science, Bharathiar University, Coimbatore, India)

Abstract: *The primary issue in designing a speech recognition system is the choice of suitable modeling unit. Speech recognition systems may be based on any one of the modeling unit like, word, phoneme and syllable. The selection of sub-word unit depends on many factors such as vocabulary size, complexity of the task, language. Phoneme is the most commonly used sub-word unit in state-of-the-art speech recognition systems, which is an indivisible unit of sound of a particular language. The choice of sub-word units, and the way in which the recognizer represents words in terms of combinations of those units, is the problem of sub-word modeling. This paper explores the various sub-word unit models used in speech recognition and presents the advantages and disadvantages of each sub-word unit.*

Keywords: *Sub-word unit, Phoneme, Syllable, Tri-phone, Acoustic segment, Speech recognition*

I. Introduction

Automatic speech recognition (ASR) is the process of converting the acoustic information in speech sequence data into its underlying linguistic structure, typically in the form of word strings [1]. As a result of serious research work in speech recognition technology, it has advanced to the level that many challenging applications are becoming a reality like voice search and interactions with mobile devices (e.g., Siri on iPhone, Bing voice search on winPhone, and Google Now on Andriod), voice control in home entertainment systems (e.g., Kinect on xBox) [2]. Some of the common speech recognition applications include dictation systems, voice user interfaces, voice dialling, call routing, domestic appliance control, command and control, voice enabled search, simple data entry, hands and eyes free applications and learning system for disabled people.

The functionality of automatic speech recognition system can be described as an extraction of a number of speech parameters from the acoustic speech signal for each word or sub-word unit. The speech parameters describe the word or sub-word by their variation over time and together they build up a pattern that characterizes the word or sub-word. In a training phase the operator will read all the words of the vocabulary of the current application. The word/sub-word unit patterns are stored and later when a word/sub-word unit is to be recognized its pattern is compared to the stored patterns and the word/sub-word unit that gives the best match is selected. It is very hard to build word models for all applications because it requires large training dataset and memory space. From the sequence of sub-word units words can be recognized with the help of sub-word unit model, pronunciation dictionary and language model. Many sub-word unit models are well applied in various speech recognition systems.

II. Modeling Units in Speech Recognition

One of the key issues in designing a speech recognition system is the selection of appropriate modeling unit for a recognition task. The choice of sub-word units, and the way in which the recognizer represents words in terms of combinations of those units, is the problem of sub-word modeling. Different sub-word models may be preferable in different settings, such as high-variability conversational speech, high-noise conditions, low-resource settings, or multilingual speech recognition. There are several recognition unit exist, such as words, phoneme, di-phone, tri-phone, senone, syllable, demi-syllable, acoustic unit, morpheme, grapheme [3] as shown in fig.1. Each of these modeling units has advantages as well as disadvantages. At a high level, the following criteria [4] need to be considered when choosing an appropriate modeling unit:

- The unit should be accurate in representing the acoustic realization in different contexts.
- The unit should be trainable where enough training data should exist to properly estimate unit parameters.
- The unit should be generalizable, so that any new word can be derived.

III. Word Models

In many traditional speech recognition systems, word models are implemented either as statistical models or templates. Words have the property that their acoustic representation is well defined and acoustic variability occurs in the beginning and end of the word. When properly trained, word models show remarkable performance in isolated and connected word recognition systems where the vocabulary size is small compared to other units. Word models are both accurate and trainable and there is no need to be generalizable. For large-vocabulary continuous speech recognition, whole word models are not suitable because of the requirement of large training data set. Each word in the vocabulary must appear several times in each context of interest. The number of words to be modeled and storage need increases abundantly. As a consequence of the above problems, the need for sub-word models thus becomes obvious.

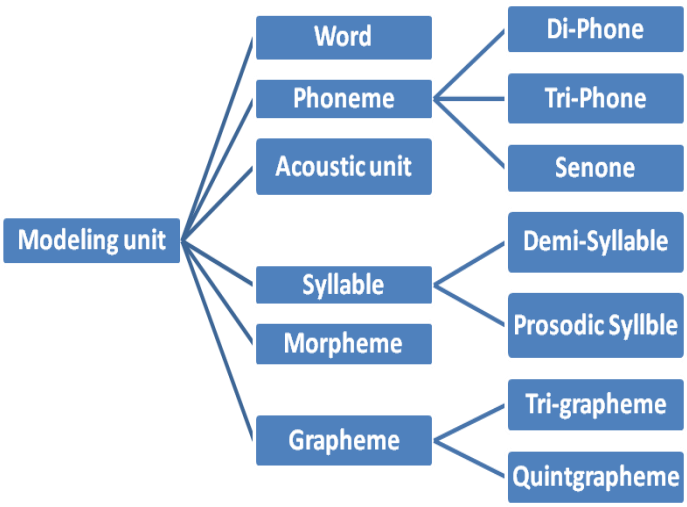

Fig.1. Modeling Units in Speech Recognition

IV. Phoneme based models

Most commonly used sub-word unit is the phone [5], a basic speech sound such as a single consonant or vowel. Each word is then represented as a sequence, or several alternative sequencesof phones specified in a pronunciation dictionary. An alternative to word models is the use of phonemes. There are typically 30-80 phones per language. For example, European languages such as English and Dutch, as well as Indian languages such as Hindi and Tamil, typically have 40 - 50 phonemes. Acoustic models based on phonemes can be trained sufficiently with as little as a few hundred sentences, satisfying the trainability criterion. Phoneme models are by default generalizable as they are the principle units all vocabulary can be constructed with. The same phone may be realized differently in different contexts, due to co-articulation, stress, and other factors. To take such effects into account, each phone in each relevant context can be considered as a separate unit. This is the context-dependent phone unit used in most speech recognizers [6]. Accuracy, however, is more of an issue, as the realization of phonemes is strongly affected by its neighboring phonemes, due to co-articulatory effects.

4.1 Phoneme Models In Foreign Languages

One of the initial works using the TIMIT database for phoneme recognition was performed by K. Lee and H. Hon [7]. It introduced mapping to 39 phonemes set for evaluation of phoneme recognition accuracy. The work used LPC-derived cepstralcoefficients, energy, delta and double-delta features. The phonemes are modeled by right-context dependent discrete HMMs and 3 codebooks of size 256 were used for basic features, deltas and double-deltas. The system used bigram language models. The results were compared with different setups with and without language models and the context-dependent system to a context-independent system and achieved 33.92 % best phoneme error rate.S.J. Young [8] took the TIMIT phoneme recognition task as an evaluation task for different approaches of HMM state tying. The input features were MFCC, log energy and deltas. The presented phoneme error rate was 38.3% which is comparably higher than the previous result but this system usedonly delta features, not double-deltas. In another work, V. V. Digalakis and colleagues [9] investigated fast search method in mono-phone stochastic segmental models. The segmental models bypass drawback of HMMs assumption of statistical independence of frames and the model attained 36% phoneme error rate.

D.J. Pepper and M.A. Clements [10] had tried to cover the whole acoustic space by one big Ergodic Discrete HMM. Then they trained another discrete HMM or a finite state automation to convert state labels to phoneme strings. The phoneme error rate (PER) obtained using discrete HMM and finite state automata are 58.5 % and 64.6 % respectively. Lamel and J. Gauvian [11] had built a phoneme recognition system using tied state Continuous Density HMMs (CDHMM). It used MFCC, log energy, deltas and double-deltas as features and trigram language model. The models were 3-state gender-dependent with tied initial and final states. The duration modeling using Gamma distribution was applied. The model achieved 26.6% phoneme error rate and comparative analysis of the model with bigram language model with other databases were presented. S. Kapadia and colleagues [12] were investigated discriminative training criterion for HMM parameter estimation. They used Maximal Mutual Information (MMI) training instead of Maximum Likelihood (ML) training. They are also compared diagonal and full covariance matrices for Gaussian Mixture Modeling. The diagonal covariance matrix system with 16 Gaussians reached 33.3 % PER using ML criterion and 32.5 % PER using MMI criterion. A full matrix system with 4 Gaussians reached 32.6 % using ML criterion and 30.7 % PER using MMI criterion.

4.2 Phoneme Models In Indian Languages

According to Census 2011, India has 122 major languages and 2371 dialects. Linguistic diversity is rich in India, out of 122 languages 22 are constitutionally recognized languages. In India there are languages like one language having many scripts and many languages having only one script. More interestingly, accent is not uniform within the same language speaking society. Irrespective of all the issues and challenges in speech recognition work, many research efforts have been made to develop high performance systems in Indian languages [13].A speaker independent, continuous speech recognition system for Hindi was proposed to recognize spoken queries in Hindi in the context of railway reservation enquiry task [14]. A spoken sentence was represented as a sequence of 48 context independent acoustic-phonetic units, each modeled by a hidden Markov model. The performance of the system for test as well as training data was collected from 10 speakers, both male and female. Sentence level accuracy and word accuracy had been analyzed and performance improvement suggestions also provided.

A small vocabulary speech recognition task in multiple Indian languages was considered in another work [15]. To configure a multi-lingual system in this task domain, an experimental study was presented using data from two linguistically similar languages – Hindi and Marathi. It is familiar that acoustic, channel and environmental mismatch between data sets from multiple languages was an issue while building multi-lingual systems of this nature. They used a cross-corpus acoustic normalization procedure which was a variant of speaker adaptive training. The resulting multi-lingual system provides the best speech recognition performance for both languages. Further, the effect of sharing similar context-dependent states from the Marathi language on the Hindi speech recognition performance was presented.Digit speech recognition is important in many applications such as automatic data entry, PIN entry, voice dialing telephone, automated banking system. A speaker independent speech recognition system for Malayalam digits was proposed [16]. The system employed MFCC as feature for signal processing and HMM for recognition. The system was trained with 21 male and female voices in the age group of 20 to 40 years and there was 98.5% word recognition accuracy on a test set of continuous digit recognition task. MohitDua et al.[17] developed isolated Punjabi word recognition system using HTK. The training carried out with 115 distinct Punjabi words and word model built and the system achieved 96% recognition accuracy. In Hindi, continuous speech recognition system to enable effective teaching geometry in primary schools was proposed which used MFCC features and HMM to create acoustic model [18]. The dataset was limited to 29 phonemes in Hindi. The Julius recognizer was used for decoding and the performance of the system tested with trained sentences and new sentences. Similarly, in Tamil language, an isolated speech recognition system to recognize grain names using HTK was developed and the model achieved better accuracy [19].

4.3 Di-Phone And Tri-Phone Models

Phonetic models can be made significantly more accurate by taking context into account, which usually refers to the immediate left and right neighboring phonemes. This leads to di-phone and tri-phone models. In phonetics, adjacent pair of phones is called di-phone. Usually, a di-phone is used to refer to the transition between two phones. If there is N number of phones in a language, then N^2 number of di-phones to be modeled. Di-phones are mainly used in speech synthesis where pre-recorded di-phones are combined to yield synthesized speech. A work was carried out by making use of context-dependent landmark-based di-phone models which require the training of both transition and internal di-phone models. Internal di-phones model the characteristics of landmarks occurring within the boundaries of a hypothesized phonetic segment, while transition di-phones model the characteristics of landmarks occurring at the boundary of two hypothesized phonetic segments [20]. A tri-phone based model takes into consideration both its left and right neighbor phone thus capturing the most important co-articulatory effects. A context may also include phones beyond immediate left or right phones

(±2). Tri-phone models decreases the word error rate by 50% compared to whole-word models and monophone models [21]. The main problem with the tri-phone models is, for the given training data set, more number of tri-phones to be modeled compared to phoneme models. This will substantially increase the computational and storage cost.Word and tri-phone based acoustic models for Tamil language was created [22]. A word based context independent acoustic model for 371 words and a Tri-phone based context dependent acoustic model for 1700 words had been built. A pronunciation dictionary with 44 base phones of Tamil and trigram based statistical language model had been created. They achieved good accuracy for trained model and test sentences read by trained and test speakers.

4.4 Senone Models

Tri-phone models ignore the similarity between tri-phones i.e. allophones. To overcome that problem similar tri-phone models can be clustered into a class called generalized tri-phones, known as parameter sharing technique [23]. Tri-phones of function words are treated as a different class because function words occur uniformly and frequent in a corpus. Another technique clusters similar states across different models that lead to finer granularity of sharing and improvement in performance. This model is called as senonemodel; technically, a senone represents a set of similar Markov states.In [24], different approaches for using deep neural networks were developed to predict senone posteriors for the task of spoken language recognition. These approaches have recently been found to outperform various baseline systems on different datasets, but still they are not compared to each other or to a common baseline. Two of these approaches used the DNNs to generate feature vectors that are then processed in different ways to predict the score of each language given a test sample. The feature extraction is made either from a bottleneck layer in the DNN or from the output layer. In the third approach, the standard i-vector extraction procedure was modified to use the senones as classes and the DNN to predict the zeroth order statistics. The three approaches were compared and the approach based on bottleneck features followed by i-vector modeling outperforms the other two approaches.Another work on modeling sub-phonetic events with Markov states and treats the state in phonetic hidden Markov models as the basic sub-phonetic unit-senone was carried out [25]. Senones generalize fenones in several ways. A word model is a concatenation of senones and senones could be shared across different word models. Senone models not only allow parameter sharing, but also enable pronunciation optimization. The authors reported preliminary senone modeling results, which have significantly reduced the WER for speaker-independent continuous speech recognition.

V. Syllable Models

Syllable is usually a larger unit than a phone, since it may encompass two or more phonemes. There are a few cases where a syllable may only consist of single phoneme. Syllables are the phonological building blocks of words. Syllables have a vital role in a language's rhythm, prosody, poetic meter and stress. The syllable as a unit, naturally accounts for the severe contextual effects among its phones as in the case of words. It accounts for pronunciation variation more systematically than a phone [26]. Syllables are longer and less context sensitive than phones and capable of exploiting both the spectral and temporal characteristics of continuous speech. Moreover, the syllable has a close connection to articulation, integrates some co-articulation phenomena, and has the potential for compact representation of conversational speech. Many ASR work carried out with syllable based models in English, Chinese and Japanese. In Japanese only 100 distinct syllables compared to English where number of distinct syllables are 30,000. When there is large number of syllables, it becomes difficult to train syllable models for ASR.

Different attempts have been made in Indian and Foreign languages to use syllables as a unit of recognition for automatic speech recognition. Some prominent work has been presented below. Ganapathiraju et al. [27] had explored techniques to accentuate the strengths of syllable-based modeling with a primary interest of integrating finite-duration modeling and monosyllabic word modeling. Wu et al. [28] tried to extract the features of speech over the syllabic duration (250ms), considering syllable length interval to be 100-250ms. Hu et al. [29] used a pronunciation dictionary of syllable-like units that are created from sequences of phones forwhich the boundary is difficult to detect. Kanokphara [30] used syllable-structure-based tri-phones as speech recognition units for Thai. A work was carried out in Amharic which is the Semitic language that has the second large number of speakers after Arabic [31]. Its writing system is syllabic with consonant-vowel syllable structure. Amharic orthography has more or less a one to one correspondence with syllabic sounds. That feature was exploited and was developed an Amharic CV syllable-based speech recognizer, using Hidden Markov Model and achieved 90% word recognition accuracy. The experiment showed that a CV syllable-based recognizer with only five emitting states performed better than all the other recognizers. However, the tri-phonebased recognizer requires much more storage space (38MB) than the syllable-based recognizer that requires only 15MB space. With regard to the speed of processing, the syllable-based model was 37% faster than tri-phonebased one.Some research work on Indian language, Tamil were carried out using syllables as sub-word unit. A novel approach to segment the speech data into syllables using group delay based processing was

done [32]. And syllable models are used to transcribe or recognize the continuous speech signal. In [33], a group delay based two level segmentation algorithm was proposed to extract accurate syllable units from the speech data. A rule based text segmentation algorithm was used to automatically annotate the text corresponding to the speech into syllable units. Isolated style syllable models are built using multiple frame size and multiple frame rates for all unique syllables by collecting examples from annotated speech. Experiments performed on Tamil language showed that the recognition performance is comparable to recognizers built using manually segmented train data. Another work in Tamil compares the performance of the syllable based models against tri-phone models [34].

5.1 Demi-Syllable And Prosodic Syllable Models

Demi-syllable based modeling approach had been tried on tonal languages [35]. According to that approach, each syllable is decomposed into two demi-syllables. The first demi-syllable does not contain tone information. The second demi-syllable, called toneme, carries the tone information of the whole syllable. In such a system, the second demi-syllable with different tones are defined as different phonemes. It leads to a successful continuous Mandarin speech recognition system. The number of basic phonetic units required to recognize tone languages is greatly reduced.It is a standard practice to decompose each Chinese syllable to a consonant initial and a final. The consonant initials do not contain tone information. Finals contain tone information. Extensive speech-recognition experiments on Cantonese and Mandarin, showed that the pitch information on the main vowel of a syllable is sufficient to determine the tone of that syllable. Based on that fact, simple and accurate speech recognition systems for tone languages can be developed which requires little training data. The method was universal, which can be applied to any tone language like Thai, Vietnamese, Japanese, Swedish, and Norwegian.In [36], prosodic syllable model was compared with syllable model in Indian language, Tamil. The results indicate that in the prosodic syllable models, there were larger number of substitution errors than that of insertions and deletions whereas in the case of word models, there was a majority of deletion errors due to morphological inflections. The overall result indicates that syllables are effective sub-word units in speech recognition.

5.2 Consonant-Vowel Models

In most of the Indian languages, consonant-vowel units occur with high frequency and those are the basic units of speech production. Hence, consonant-vowel units can be used as recognition unit and effective models can be designed for Indian languages. Some of the CV unit recognition work carried out in Indian languages are studied in detail and elaborated as follows. Neural network models were developed for spotting Stop-Consonant-Vowel (SCV) segments in continuous speech [37]. Gangashetty et al. proposed CV unit recognition for isolated Hindi utterances using Auto Associated Neural Networks (AANN) [38]. CV classes are sub-grouped and different classifiers are modeled based on Manner of Articulation (MOA), Place of Articulation (POA) and vowels. The work was extended [39] for spotting multilingual CV units in speech using neural network models. They used frequently occurring CV units of three Indian languages Tamil, Telugu and Hindi.

In another work, Vuppalaet.al, [40] experimented CV unit recognition task on Telugu speech data under background noise using temporal and spectral processing methods. Recognition of vowel category of CV unit and then consonant category of CV unit is done separately and SVM, HMM hybrid models are employed to improve the performance. Thasleema et.al,[41] carried out Malayalam CV unit classification using different classifiers like support vector machine with Decision Directed Acyclic Graph learning architecture (DDAGSVM), K-Nearest Neighbour and Artificial Neural Network. They classified 5 classes of CV units which are unaspirated, aspirated, approximants, nasals and fricatives using hybrid wavelet based features which work better on noisy speech data. AnanthaNatarajan et.al,[42] proposed an approach in which the continuous speech is segmented into smaller speech units and each unit is classified either consonant or vowel using the formant frequencies on Tamil broadcast data.

VI. Morpheme Based Models

Out-of-vocabulary (OOV) words are a major source of error in a speech recognition system and various methods have been proposed to increase the performance of the systems by properly dealing with them. Languages which are morphologically rich can use morphemes as modeling unit for speech recognition. In such as way, the work on developing speech recognition system for Amharic, a morphologically rich language was carried out [43]. To solve the OOV problem by using morphemes as dictionary and language model units was tried. It has been found that for a small vocabulary system morphemes are better lexical and language modeling units than words. An absolute improvement in word recognition accuracy of 11.57% was obtained as a result of using a morpheme based vocabulary. However, for large vocabularies morphemebased systems did not bring much performance improvement as they suffer from acoustic confusability and limited language model scope.

An attempt was made in Korean large vocabulary continuous speech recognition with morpheme-based recognition units [44]. In Korean writing, a space is placed between two adjacent word-phrases, each of which generally corresponds to two or three words in English in a semantic sense. If the word-phrase is used as a recognition unit for Korean large vocabulary continuous speech recognition (LVCSR), the out-of-vocabulary (OOV) rate becomes very large. If a morpheme or a syllable is used instead, a severe inter-morpheme co-articulation problem arises due to short morphemes.A merged morpheme as the recognition unit and pronunciation-dependent entries in a language model was proposed to reduce such difficulties and incorporate the between-word phonology rule into the decoding algorithm of a Korean LVCSR system. Starting from the original morpheme units defined in the Korean morphology, pairs of short and frequent morphemes are merged into larger units by using a rule-based method and a statistical method. The merged morpheme unit was defined as word and used it as the recognition unit. The performance of the system was evaluated in two business-related tasks: a read speech recognition task and a broadcast news transcription task. The OOV rate was reduced to a level comparable to that of American English in both tasks.

VII. Grapheme Based Models

In some phonetic languages the relation between written and spoken form is reasonably close which leads to successful grapheme based speech recognizers (GBSR). Many research works has proved that, instead ofphonemes, graphemes – the orthographic representation of a word can be used as the sub word units. The main challenge in grapheme based speech recognition task, generation of the pronunciation dictionary. One of essential components in the process of building speech recognizer is pronunciation dictionary that maps orthographic representation into a sequence of phonemes — the sub words units, which we use to define acoustic models during the process of training and recognition. The acquisition of quality hand-crafted dictionary requires linguistic knowledge about target languages and is time- and money-consuming, especially for rare and low-resource languages. The most straightforward method is to generate pronunciation dictionary as sequence of graphemes and thus to directly use orthographic units as acoustic models.

GBSR systems have been successfully applied to several European languages. A model was developed in three languages - English, German, and Spanish - were trained and compared to their phoneme based counterparts. The experiments were carried out on GlobalPhone speech corpus that provides clean read speech in 15 different languages. Initially monolingual phoneme based recognized for all three languages developed. The same database, preprocessing, HMM-architecture, and language model as their phoneme based counterparts used for developing the grapheme based recognizers. The only difference lies within the subunits, the pronunciation dictionary, and the question set for creating the context dependent models. Pronunciation dictionaries for the grapheme based recognizers are built by simply splitting a word into its graphemes.

Graphemes with diacritics like the German umlaut $u\ddot{}$ were treated as an independent grapheme. Digits and numbers were preprocessed by rule-based digit-to-grapheme scripts. As in the case of phonemes, a grapheme is modeled by a 3 −state HMM consisting of a begin, a middle, and an end-state. The Spanish and German results imply that the grapheme based approach is feasible for languages with a good grapheme-to phoneme relation. However, for English with its fairly poor grapheme-phoneme correspondence the grapheme based system is significantly outperformed by the phoneme based one [45]. Additionally, multilingual grapheme based recognizers are designed to investigate whether grapheme based information can be successfully shared among languages.A grapheme-based speech recognition system for Thai was developed with Thai GlobalPhone corpus and different settings for the initial context independent system, different number of acoustic models and different contexts for the speech unit [46]. An enhanced tree clustering method was used as a way of sharing parameters across models. The system had been compared with phoneme-based systems built with hand-crafted dictionary and automatically generated dictionary. Experiment results show that the grapheme-based system with enhanced tree clustering performs better than the phoneme-based system using an automatically generated dictionary, and has comparable results to the phoneme-based system with the handcrafted dictionary. They could reduce the word error rate upto 26% in Trigrapheme model and 30% in Quintgrapheme models.

The experiments are extended for some other languages and proved that grapheme-based speech recognition that manages with the problem of low-quality or missing pronunciation dictionaries, is applicable for phonetic languages and tonal languages like Vietnamese. For non-phonetic languages, like English, using of models with wider context provides comparable results and grapheme based approach can be, with small limitation, usable also for this class of non-phonetic languages. This direct approach, supported by the expansion of numbers in dictionaries, is beneficial particularly in case of low-resource languages and could be successfully used in building speech recognizers for rare languages [47].A novel grapheme-based ASR system was developed for English where the correspondence between phoneme and grapheme is weak. The system jointly modeled phoneme and grapheme information using Kullback-Leibler divergence-based HMM system (KL-HMM). They were able to achieve comparable results for trigrapheme and quintgrapheme models over phoneme based models [48].

Another work was proposed on grapheme-based ASR in the framework of Kullback-Leibler divergence based hidden Markov model (KL-HMM) for under-resourced languages, particularly Scottish Gaelic which has no lexical resources. Scottish Gaelic is one of three primary Goidelic languages. Classified within the Indo-European language family, it is limited within the group of Celtic languages, and as such is only distantly related to any of the well-resourced major European languages. The results showed that, all KL-HMM systems yield better performance than HMM/GMM system for both mono and tri contexts [49].

VIII. Conclusion

The effective design and development of speech recognition system depends upon the selection of suitable modeling unit. Speech recognition systems may be based on any one of the recognition unit such as word, phoneme, syllable, morpheme and grapheme. This paper presented the extensive study on different sub-word unit models used in speech recognition systems. It also elaborates selection of sub-word units highly dependent on the nature of language.

References

[1]. Li Deng, Xiao Li, Machine Learning Paradigms for Speech Recognition: An Overview, IEEE Transactions on Audio, Speech, and Language Processing, 21(5), 2013, 1060-1089.

[2]. Jinyu Li, Li Deng, Reinhold Haeb-Umbach, Yifan Gong, "Robust Automatic Speech Recognition: A Bridge to Practical Applications", Academic Press, 2015.

[3]. L. R. Rabiner, B. H. Juang, B. Yegnanarayana, "Fundamental of Speech Recognition", Pearson Education Inc., New Delhi, India, 2009

[4]. K. Livescu, E. Fosler-Lussier, F. Metze, "Sub-word modeling for automatic speech recognition: Past, present, and emerging approaches", IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 44–57, 2012

[5]. B. H. Repp,"On levels of description in speech research", The Journal of the Acoustical Society of America, vol. 69, no. 5, pp.1462–1464, 1981.

[6]. J. Odell, The Use of Context in Large Vocabulary Speech Recognition, Ph.D. thesis, University of Cambridge, March 1995.

[7]. K. Lee, H. Hon, "Speaker-independent phone recognition using hidden markov models",IEEE Transactions on Acoustic, Speech and Signal Processing, vol. 37, no. 11, pp. 1641–1648, Nov. 1989.

[8]. S. J. Young, "The general use of tying in phoneme-based hmm speech recognizers," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP), San Francisco, USA,Mar 1992.

[9]. V. V. Digalakis, M. Ostendorf, J. R. Rohlicek, "Fast algorithms for phone classification and recognition using segment-based models," in IEEE Transactions on Signal Processing, Dec 1992, vol. 40, pp. 2885–2896.

[10]. D. J. Pepper and M.A. Clements, "Phonemic recognition using a large hidden markov model", IEEE Transactions on Signal Processing, vol. 40, no. 6, pp. 1590 – 1595, Jun 1992.

[11]. L. Lamel, J. Gauvian, "High performance speaker-independent phone recognition using cdhmm", in Proc. European Conf. Speech Communication and Technology, 1993, pp. 121–124.

[12]. S. Kapadia, V. Valtchev, S. J. Young, "MMI Training for Continuous Phoneme Recognition on the TIMIT Database. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP '93), Minneapolis, vol. 2, 1993, pp. 491–494.

[13]. Mousmita Sarma, Kandarpa Kumar Sarma, "Recent Trends in Intelligent and Emerging Systems",Part of the series Signals and Communication Technology, Springer India, 2015, pp. 173-187.

[14]. Samudravijaya K, "Computer Recognition of Spoken Hindi", Proc. Int. Conf. Speech, Music and Allied Signal Processing, Thiruvananthapuram, Dec. 2000, pp 8-13.

[15]. Mohan A, Rose R, Ghalehjegh SH, Umesh S, "Acoustic modelling for speech recognition in Indian languages in an agricultural commodities task domain". Speech Communication, 56 (2014), pp. 167–180.

[16]. Cini Kurian, Kannan Balakrishnan,"Speech Recognition of Malayalam Numbers", IEEE Transaction on Nature and Biologically Inspired computing, (NaBIC-2009), pp 1475-1479.

[17]. Mohit Dua, R.K.Aggarwal, Virender Kadyan, Shelza Dua, "Punjabi Automatic Speech Recognition Using HTK", International Journal of Computer Science Issues, Vol. 9, Issue 4, July 2012

[18]. Gaurav, Devanesamoni Shakina Deiv, Gopal Krishna Sharma, Mahua Bhattacharya, "Development of Application Specific Continuous Speech Recognition System in Hindi", Journal of Signal and Information Processing, 3, 2012, pp. 394-401.

[19]. Akila A, Chandra E, "Isolated Tamil Word speech Recognition System using HTK", International Journal of Computer Science Research and Application, vol. 3, Issue 2, pp 30-38, 2013.

[20]. James R Glass, Timothy J. Hazen, I. Lee Hetherington,"Real-time telephone-based speech recognition in the Jupiter domain", ICASSP 1999, pp. 61-64.

[21]. Bahl L.R., Bakis R., et.al, "Further results on the recognition of a continuously read natural corpus", IEEE international conference on acoustics, Speech and Signal Processing,1980.

[22]. Thangarajan R, Natarajan AM, Selvam M,"Word and Triphone Based Approaches in Continuous speech Recognition for Tamil Language", WSEAS Transactions on Signal processing, Vol 4, 3, pp 76-85, 2008

[23]. Hwang M.Y, Huang X.D.,"Shared distribution hidden Markov models for speech recognition", IEEE Transactions on Speech and Audio Processing, vol. 1, No.4, pp.414-420.

[24]. L. Ferrer, Y. Lei,M. McLaren, "Study of senone-based deep neural network approaches for spoken language recognition", IEEE Trans. Audio Speech and Language Processing, Issue 99, 2015, pp.1-1.

[25]. M. Hwang, X. Huang, "Subphonetic Modeling with Markov States – Senone", Proc. IEEE ICASSP-92, 1:33–36, San Francisco, CA, March 1992.

[26]. Greenberg. S, "Speaking in shorthand-A syllable-centric perspective for understanding pronunciation variation", Speech Communication 29 (1999), pp. 159-176.

[27]. A Ganapathiraju, V Goel, J Picone, A Corrada, G Doddington, K Kirchhoff, M Ordowski, B Wheatley, "Syllable – a promising recognition unit for LVCSR," in Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, 1997

[28]. Su-Lin Wu, Brian Kingsbury, Nelson Morgan,and Steven Greenb, "Information from Syllable-length Time Scales into Automatic Speech Recognition", ICASSP-98, Seattle, pp. 721-724.

[29]. Hu, Z., Schalkwyk, J., Barnard, E., & Cole, R., " Speech recognition using syllable-like units. In Proceedings of the International Conference on Spoken Language Processing, Philadelphia, USA, October 1996, vol.2, pp.1117-1120.
[30]. Supphanat Kanokphara et al., "syllable structure based phonetic units for context-dependent continuous Thai speech recognition", European conference on speech communication and Technology, 2003, pp. 797-800.
[31]. Solomon Teferra Abate, Wolfgang Menzel, "Syllable-Based Speech Recognition for Amharic", Proceedings of the 5th Workshop on Important Unresolved Matters, pages 33– 40,Prague, Czech Republic, June 2007.
[32]. T.Nagarajan, Hema A.Murthy and Rajesh M.Hegde, "Segmentation speech into syllable like units", EUROSPEECH-2003, pp. 2893- 2896.
[33]. Lakshmi. A, Hema A. Murthy, "A Syllable based continuous speech recognizer for Tamil", SPCOM, 2004.
[34]. R. Thangarajan, A. M. Natarajan, M. Selvam, "Syllable modeling in continuous speech recognition for Tamil language", International Journal of Speech Technology, 2009, Volume 12, Number 1, pp.47-57.
[35]. C. Julian Chen, Haiping Li, Li Qin Shen, Guokang Fu, "Recognize tone languages using pitch information on the main vowel of each syllable", In proc. ICASSP 2001, pp. 61-64.
[36]. R. Thangarajan, "Speech Recognition for Agglutinative Languages", Modern Speech Recognition Approaches with Case Studies, Dr. S Ramakrishnan (Ed.), InTech, 2012.
[37]. C. C. Sekhar and B. Yegnanarayana, "Recognition of stop-consonant-vowel (SCV) segments in continuous speech using neural network models," Journal of the Institution of Electronics and Telecommunication Engineers, Vol. 42, Issue 4 and 5, July-October. 1996, pp. 269-280.
[38]. S. V. Gangashetty, C. C. Sekhar, and B. Yegnanarayana, "Detection of vowel onset points in continuous speech using autoassociative neural network models," In Proceedings of International conference on spoken language processing, Korea, October 2004, pp. 1081-1084.
[39]. S. V. Gangashetty, C. C. Sekhar, and B. Yegnanarayana, "Spotting multilingual consonant-vowel units of speech using neural network models," ISCA Tutorial and Research Workshop on Nonlinear Speech Processing, Barcelona, pp. 287-297, April 2005.
[40]. A. K. Vuppala, K. Sreenivasa Rao and Saswat Chakrabarti, "Improved consonant-vowel recognition for low bit-rate coded speech," Wiley International Journal of Adaptive Control and Signal Processing, Vol. 26, Issue 4, pp. 333-349, 2011.
[41]. T. M. Thasleema and N. K. Narayanan, "Wavelet transform based consonant - vowel (CV) classification using SVMs," ICONIP 2012, Part II, LNCS 7664, Springer, pp. 250-257, 2012.
[42]. V. Anantha Natarajan and S. Jothilakshmi, "Segmentation of continuous speech into consonant and vowel units using formant frequencies," International Journal of Computer applications, Vol. 56, Issue. 15, October 2012.
[43]. M. Y. Tachbelie, S. T. Abate, and W. Menzel, "Morpheme-based automatic speech recognition for a morphologically rich language - amharic," in SLTU'10, 2011.
[44]. Oh-Wook Kwon, Jun Park, "Korean large vocabulary continuous speech recognition with morpheme-based recognition units", Speech Communication, Vol. 39, No.3-4, 2003, Pages 287–300.
[45]. Mirjam Killer, Sebastian Stuker, Tanja Schultz "Grapheme based Speech Recognition", Interspeech 2003.
[46]. Paisarn Charoenpornsawat, Sanjika Hewavitharana, Tanja Schultz, "Thai Grapheme-Based Speech Recognition", In Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume, pages 17–20,2006.
[47]. M. Janda, "Grapheme Based Speech Recognition", In Proceeding of the 18th Conference STUDENT EEICT 2012, Brno, CZ, vol. 3, pp. 441-445, April 2012.
[48]. Mathew Magimai.-Doss, Ramya Rasipuram, Guillermo Aradilla, Herve Bourlard, "Grapheme based automatic speech recognition", Interspeech 2011.
[49]. Ramya Rasipuram, Peter Bell and Mathew Magimai.-Doss, "Grapheme and Multilingual Posterior Features for Under-Resourced Speech Recognition: A Study on Scottish Gaelic", In Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing, May 2013.