ISSN 2319-2720

Volume 2, No.1, January – March 2013
International Journal of Computing, Communications and Networking
Available Online at http://warse.org/pdfs/2013/ijccn01212013.pdf

Word and Syllable Boundary of Sylheti Phonemes/ Syllables

Suchismita Sinha[1], P.H.Talukdar[2]
Dept. of Instrumentation & USIC, Gauhati University, Assam.
[1]shelleyarjunsil@gmail.com
[2]phtassam@gmail.com

ABSTRACT
It is the aim of this paper to explore the various pronunciation lexicon features of Sylheti language. Presently, we are attempting to identify the basic syllable features related to a word and then the sentences through the pitch and Intensity analysis approaches. Since Sylheti is very fast language it is expected that this findings may be very useful while designing any TTS or speech synthesizers for Sylheti language.

Key words: pronunciation lexicon, syllable, pitch, Intensity.

1. INTRODUCTION

A word is the smallest free form (an item that may be uttered in isolation with semantic or pragmatic content) in a language, in contrast to a morpheme, which is the smallest unit of meaning. A word may consist of only one morpheme (e.g. so, very), but a single morpheme may not be able to exist as a free form (e.g. the English plural morpheme -s). Typically, a word will include a root or stem, and it may also include one or more affixes. Words can be combined to create other units of language, such as phrases, clauses, and/or sentences. A word consisting of two or more stems joined together form a compound.

Word may refer to a spoken word or a written word, or sometimes, the abstract concept behind either. Spoken words are made up of phonemes. Words are combined to create phrases, clauses, and/or sentences. A compound is formed joining two or more word(s). In a spoken language, the distinction of individual words is usually given by rhythm or accent. In a synthetic language, a single word stem may have a number of different forms. In these languages words are constructed from a number of morphemes. For example, in the Indo-European languages, the morphemes are distinguished through the use of optical suffixes, orthography and word. In modern orthography of language, word separators (typically spaces) are common syntactic languages which often combine lexical morphemes into single words. In polysynthetic languages, script use single character to represent a word. But most existing scripts are partly logographic and partly combination of both logographic and phonetic units.

Unlike English and other western languages, Asian languages such as Chinese, Japanese, Thai etc. there are no space boundaries between words. This posts a problem where translating these languages into English. When listen to speech, we hear a sequence of words, but when we speak, we do not separate-words-by-pause. A first step to learn the words of a language, and then is to extract the words from continuous speech. Remarkably, 7.5 month-old infants are reported extracting words from speech [1]. The problem of word segmentation has been one of the most important research areas in developmental psychology. It appears that the problem of word segmentation would be simple and easier if all utterances consist of only isolated word. The problem of word segmentation is particularly significant in the parsing of written text in languages that do not explicitly include spaces between words. Children, with little or no knowledge of the inventory of words a language possesses, the identification of word boundaries by a child is a significant problem in the domain of child language acquisition [2]. The segmentation of the continuous speech into isolated words seems easy to native speaker. But the difficulty of the task can be realized when one listens to a language not known to him. Often, adults listening to a foreign language remark that the speakers of that language speak rapidly. This difficulty in segmenting words occurs because unlike text, whose every word is separated by blank spaces, in speech no prominent marking is present between words. Different languages require different cues for word segmentation as they vary in:-

- Prosody
- Phonetics,
- Phonotactics and
- Other distributional Properties.

There are number of potential sources of information that could be used as indicators of word boundaries influent speech. These include:

(i) Metrical stress cues,
(ii) Phonotacties cues,
(iii) Context-sensitive allophones
(iv) Co-articulation cues and
(v) Statistical/distributional properties.

Infants are generally more bias towards troches i.e. to listen words with strong/weak stress pattern i.e. words with a weak / strong pattern [3]. This may be a strong evidence of the usage of metrical stress cues by infants. Another important cues in the word segmentation problem - the co-occurrence relations among syllables. The child learning language is faced with a dounting task, for example, extracting meaning from an apparently meaningless stream of sound. To solve this problem, a number of contradictions and confusion must be resolved. These contradictions and confusion ranges from segmenting individual words out of the acoustic stream to understanding. Behavioral research is the usual method employed to study the language acquistion, but due to complex interdependence of different components of language, it is difficult to determine exactly how each component contributes to the overall learning process. Language acquisition is a problem of induction -the creation of an internal representation of language that allows learners to generalize beyond the observed linguistic input, interpreting and producing novel lingusitcs forms. Two standard theoretical explanations are:

- Nativism
- Empiricism.

Some questions about the nature of lexical acquisition must be resolved:

(i) What kinds of structures are to be considered by the learning mechanism including both child and machine?
(ii) How much and what sort of evidence is necessary to produce generalization?
(iii) Are these innate constraints that are specific to language acquisition, or can language be acquired successfully using only general learning biases?
(iv) What kinds of interactions between linguistic components aid in learning?

Allophonic cues are also an important kind of potential cues. It is true that certain allophonic variants of phonemes occur in certain positions in the words. Thus, there are many different kinds of cues predicted, used for the determination of word boundaries, each targeting a different aspects of speech and useful in its own respect with insufficient independency. A word is not learned until it becomes a part of the lexicon. The sound patterns of words may be extracted and stored independently prior to learning of word meanings. Jusczyk and Hohne, 1997[2] reported that – " 8-month old infants were familiarized with novel words embedded in stories with speakers as well as order variations. Two week later, they listen to the previously heard words significantly longer than foil words with similar phonetic characteristics and overall frequency to which they had not been exposed. This apparently takes place well before the learning of word meanings." This assumed a lot of importance when we think of the big picture, i.e. using the word boundary recognition system as a part of a speech recognition system. Thus, it is almost an established fact that children workout where the word boundaries are through:

- Pauses(although this is dubious),
- Intonation(this too is dubious), and
- Statistical regularities.

Among the other phonological cues that may help segmenting the speech stream are:

- Voiceless stops that begin words are almost always aspirated.
- Voice segments that end words are often de-voiced.
- Various other phonological processes may occur i.e. word-final frication etc

The statistically related aspects of word boundary prediction are mostly applicable to phonologically related things. Statistical tracking is only useful for auditory stimuli, and not visual. Apart from human, Vervet and Tamarin monkeys have been shown to have essentially the same abilities that humans do. Statistically this is an established fact that the transition probability is high within a word, but low across a word boundary.

2. NATIVISM

It is assumed in theories that general learning mechanisms simply are not powerful enough. .Mostly, children acquire first or second language quite successfully even though no special care is taken to teach them and no special attention is given to their progress. It also seems apparent that much of the actual speeches consist of fragments and deviant expressions of a variety of sorts. Thus, a child must love the ability to "invent" a generative grammar that defines well-form and assigns interpretations to sentences. In other words, the children are not explicitly thought of language, and because their linguistics input is noisy, they must not be learning language entirely from the input, but rather "inventing" it is some sense. Many nativists have bolstered this claim by arguing that– language is un-learnable even in absence of noise [4]. In practice, nativists generally assume that linguistics representations are highly structured, consisting of categories, rules and the like [5]. On the other hand, psychologists and linguists who disagree with the nativist approach often align themselves instead with the empiricist view of language acquisition. They proposed that– language acquisition is based on statistical properties of the input, and is an associative process. Thus, where the nativists see children's input as noisy with a lacking of full complexity

of adult language, the empiricists see the same with rich statistical features. They are more sceptical of language-specific mental representations. On the other hand, the connectionists advocate it as probabilities, distributed and constructured, without any hard categorical boundaries or explicit rules [6]. They view that linguistic rules and categories have no cognitive reality. Overall, many of the difference between empiricism and nativism is framed as differences in the inductive bias of the learner– the strength and nature of constraints on learning. Nativists think the learning as highly constrained by the nature of linguistic representations and assume that their representations can be canonical. Empiricism argue the constraints as weak and the learning is guided by the nature of input. Although speech lacks explicit demarcation of word boundaries, it possesses some significant cues, which can help or lead us to ascertain the word boundaries correctly.

3. OBJECTIVE

Sylthei is one of the oldest language covering the present Bangladesh and Southern part of North-East region of India. During the period of British rule and Undivided India this language was very popular and major link language in this part of Asian sub-continent. After the British rule is over and Independence of the East Pakistan (renamed as Bangladesh) this language gradually confined into a limited domain due to various socio-political reasons. As a result the original and native speakers of this language come across many omission and addition while exchanging their views and ideas with other major communities of this region. Surprisingly, the original scripts are also either distorted or rejected. At this point of time to justify the accountability of this century old language a detail study on the various issues related to this language is a must.

It is the aim of this paper to explore the various pronunciation lexicon features of this language. Presently, we are attempting to identify the basic syllable features related to a word through the pitch and Intensity analysis approaches. These features would be used further to develop machine and speaker independent complete speech synthesis system.

4. WORD SEGMENTATION

The submitting author is responsible for obtaining agreement of all coauthors and any consent required from sponsors before submitting a paper. It is the obligation of the authors to cite relevant prior work.

Various methods have been proposed to address the problem of word segmentation. They can be classified into three categories:

- Purely dictionary based approach
- Purely statistical based approach, and
- Statistical based approach using manual segmentation data.

The purely dictionary based approach follow a matching hierarchy. Based on a given word set the word boundaries are detected through an algorithm. The algorithm searches a sentence from left to right for the longest sequence of characters that match a word in the dictionary and insert the boundary to that point. This matching heuristic is very simple to implement. But its performance is completely dependent on the coverage of the dictionary and fails to detect boundaries for words not in the dictionary.

The statistical- based approach is based on the mutual information (or transition frequency) of adjacent characters to detect the word boundaries. Group of characters having mutual information greater or higher than a threshold, following the given sequence of words, form a word. In this approach, there is no requirement of an external dictionary. It can be applied to any languages. However, though more logical and realistic, this approach does not perform as well in terms of segmentation accuracy. The statistical - based approach using manual segmentation data is found suitable for languages which do not have any space between words. Here, the segmentation of words is basically done using tags i.e. character tags. These tags are assigned at the beginning of a new word and/or in the middle and/or at the end of the word. The task of assigning a sequence of tags to a sentence is somewhat contradictory in this approach. However, this approach enjoyed some score of acceptability in the ambiguity resolution and unknown word detection. Thus, a number of different cues to word boundaries are present in fluent speech. Infants are able to use many of these including phonotactics, allophonic variation [7], metrical (stress) patterns, effect of co-articulation [8], and statistical regularities amongst sequences of syllables [9]. Most work on statistical word segmentation is based on the transition probabilities. This has obviously lead us to the fact that infants use statistics such as mutual information or transition probabilities while segmenting words from speech. All the statistical observations about the predictability of word boundary are consistent with two basic assumptions, either a word is a unit that is statistically independent of other units, or it is a unit that helps to predict other units. The ease or difficulty of deciphering a word depends on the language. Dictionaries categorize a language's lexicon (i.e., its vocabulary) into lemmas. These can be taken as an indication of what constitutes a "word" in the opinion of the writers of that language.

5. WORD BOUNDARIES

The task of defining what constitutes a "word" involves determining where one word ends and another word begins—in other words, identifying word boundaries. There are several ways to determine where the word boundaries of spoken language should be placed:

- **Potential pause:** A speaker is told to repeat a given sentence slowly, allowing for pauses. The speaker will tend to insert pauses at the word boundaries. However, this method is not foolproof: the speaker could easily break up polysyllabic words, or fail to separate two or more closely related words.
- **Indivisibility:** A speaker is told to say a sentence out loud, and then is told to say the sentence again with extra words added to it. Thus, I have lived in this village for ten years might become My family and I have lived in this little village for about ten or so years. These extra words will tend to be added in the word boundaries of the original sentence. However, some languages have infixes, which are put inside a word. Similarly, some have separable affixes.
- **Phonetic boundaries:** Some languages have particular rules of pronunciation that make it easy to spot where a word boundary should be. For example, in a language that regularly stresses the last syllable of a word, a word boundary is likely to fall after each stressed syllable. Another example can be seen in a language that has vowel harmony (like Turkish):]the vowels within a given word share the same quality, so a word boundary is likely to occur whenever the vowel quality changes. Nevertheless, not all languages have such convenient phonetic rules, and even those that do present the occasional exceptions.

In practice, linguists apply a mixture of all these methods to determine the word boundaries of any given sentence. Even with the careful application of these methods, the exact definition of a word is often still very elusive.

In the present study I have applied a mixture of all the above mentioned processes to the Sylheti sentences uttered by Sylheti speakers both male and female taken randomly from the Sylheti speaking area. The sentences have been recorded using Cool Edit Pro.2 and sampled at 20500 Hz and I have analysed the sentences using the software Praat. The plots of the syllable and word boundary analysis of the Sylheti utterances are given in Figure 1 to Figure 4 where the yellow curve shows intensity and the red curve shows pitch. The utterances are typical Sylheti sentences which are as follows:

- AMI GORO JAITAM
- OU DEKHI OU NAI

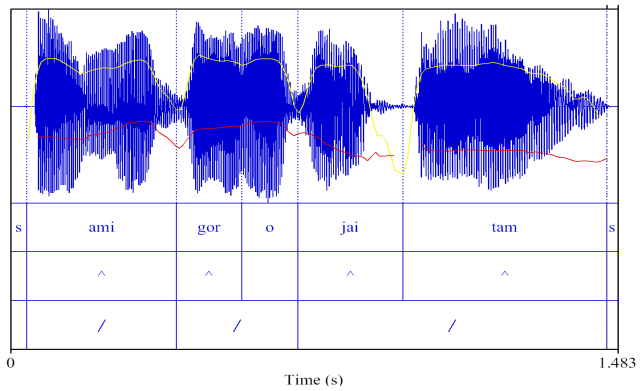

Figure 1: Plots of the syllable and word boundary analysis of female utterances of AMI GORO JAITAM.

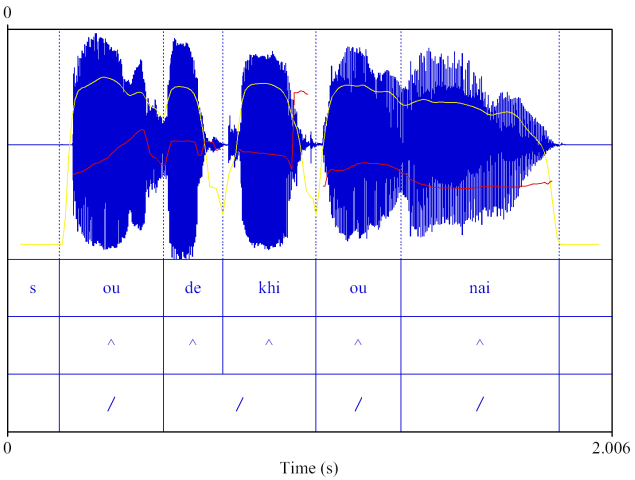

Figure 2: Plots of the syllable and word boundary analysis of male utterances of AMI GORO JAITAM.

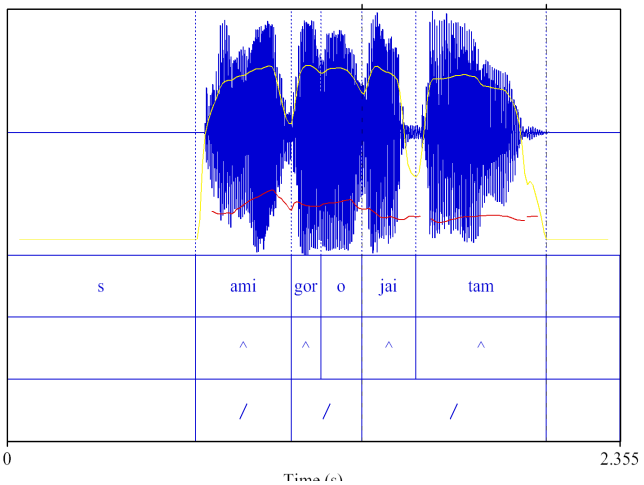

Figure 3: Plots of the syllable and word boundary analysis of female utterances of OU DEKHI OU NAI.

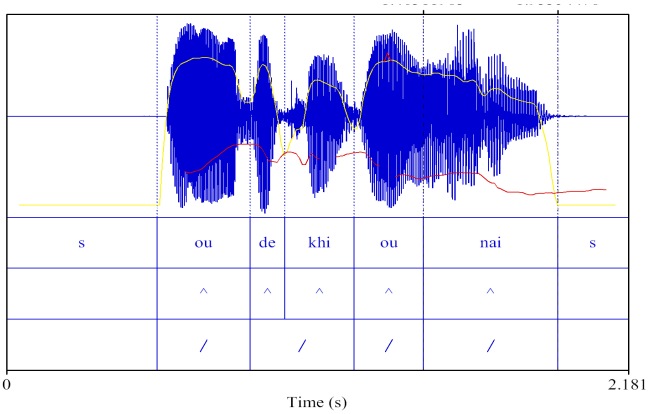

Figure 4: Plots of the syllable and word boundary analysis of male utterances of OU DEKHI OU NAI.

5. OBSERVATION AND CONCLUSION

As shown in the figures Fig 6.1(a), (b) it is found in the present word/syllable boundary analysis of Sylheti language that the measure of intensity and pitch play a significant role in determining the word / syllable boundary. As in the case of earlier reports for Assamese [10], Bengali [11], Hindi [12], in case of Sylheti language also, through intensity and pitch analysis we can determine the syllable and word boundary from speech spectra.

The marked depression of the intensity measure between two words is prominent in spectras. However, the depression is less in case of syllable boundary, but yet distinguishable. Similarly, the pitch curve shows the syllable boundary and word boundary either through fall of pitch or breaking of pitch continuity. Thus, through the pitch and intensity measure we can effectively locate the boundary of syllable and word in Sylheti language. Since Sylheti is a very fast speaking language, these informations may help in the development of the Speech–To-Text and Text-To-Speech systems, and speech synthesisers, for Sylheti language.

REFERENCES

1. Jusczyk, P. & Aslin, R. **Infant's detection of the source patterns of words fluent speech**, cognitive psychology, Vol. 29, pp 1-23, 1995.
2. Jusczyk, Peter, W and E.A. Hohne. **Infant's memory for spoken words, science**, Vol. 277 no. 5334 pp. 1984-1986 ,1997.
3. Jusczyk, P.W. Cutler, A. & Redanz, N. **Preference for the predominant stress patterns of English Words.** Child Development, Vol. 64, no. 3, pp. 675-687, 1993.
4. E.M. Gold. **Language identification in the limit.** Information and Control, Vol. 10, pp. 447-474 , 1967
5. Pinker, S. **Language learn ability and language development,** Cambridge, M.A. : MIT Press, 1984.
6. Elman, J.L. **Finding structure in time**, cognitive science, Vol. 14, pp. 179-211, 1990.
7. Jusczyk, P.W. **Narrowing the distance to language: one step at a time**, J. Comon Disord, Vol. 32, pp. 207-222, 1999.
8. Johnson, E.K. & Jusczyk, P.W. **Word segmentation by –month olds: when speech cues can't be more than statistics**. J. of memory and language, Vol. 44, pp. 548-567, 2001.
9. Saffron, J.R., Newport, E.L. and Aslin, R.N. **Word segmentation: The role of distributional cues**. J. of memory and language, Vol. 35, pp. 606-621, 1996.
10. Bodo M.R, Assamese and Bodo, **A comparative and contrastive study**, Priyadini Publication, 1990, 1st Edition.
11. Dash, N.S and B.B. Choudhury, **A Corpus- Based Study Of The Bangla Language**, Indian Journal of Linguistics, 2001.
12. Sunita Arora, Karunesh Kr. Arora, S.S Agarwal, **Statistical Text Analyzer for Hindi And Other Indian Languages**,Workshop on Spoken Language Processing, TIFR, Mumbai,2003.