ISCA Archive
http://www.isca-speech.org/archive

Speech and Language Technology in Education (SLaTE2011)
Venice, Italy
August 24-26, 2011

A Vocabulary Acquisition Framework using Audio Books

Meena Vundela, Swetha A.V.S.L.G and Kishore Prahallad

International Institute of Information Technology, Hyderabad, India.
{meena.vundela,swetha.avslg}@students.iiit.ac.in, kishore@iiit.ac.in

Abstract

In this paper, we describe a framework for vocabulary acquisition using audio books. The proposed framework performs automatic alignment of text and large speech files in audio books. This enables highlighting of words and sentences during voice over. The proposed framework can be easily extended to a large number of audio books. We evaluate this framework on a set of subjects between the age of 7-13 years, and demonstrate the effectiveness of such framework for vocabulary acquisition.
Index Terms: Audio books, Incidental vocabulary acquisition

1. Introduction

Acquisition of words and their meanings is an important task in learning a language. The process of acquisition of vocabulary of a language can be incidental or intentional. An example of incidental acquisition is learning the vocabulary through participation and conversation in a day-to-day activity. Intentional learning is often encountered when a child or an adult is learning a second language. Earlier techniques for intentional learning include – 1) flash cards promoting intentional memorization, 2) use of pictures and the associated description such as Rosetta Stone (http://www.rosettastone.com), 3) retrieval of appropriate reading material from the Internet [1] and 4) generation of a pseudo-sentence/phrase for practice [2]. Techniques for incidental acquisition include – 1) speech enabled card games [3] and 2) task based vocabulary learning [4]. As noted in [3], there are few works which provide an environment/framework for incidental vocabulary acquisition. In this work, we describe a story-telling approach for incidental vocabulary acquisition using audio books. The idea is to expose learners to short stories with voice over. At any given time, the learner/user can point to a word, to have its meaning highlighted. The use of audio books not only attracts the attention of the learners, but also provides enough motivation for language acquisition at various levels including vocabulary, pronunciation, grammar and intonation.

It is well known in the educational community that children understand and remember the words better when they are explained the meanings based on usage of words, especially in the context of a story [5]. The scope of this work is to build an automated framework for vocabulary acquisition using audio books. We refer to this framework as VAA (vocabulary acquisition using audio books). Using this framework, the learners will get to know the meanings as they run through story. This framework employs an automatic tool for aligning text with large speech files in audio books, and thus it can be easily extended to a large number of audio books available in public domain. The learning gain of vocabulary acquisition is tested on local children in Hyderabad, India. As far as our knowledge, such a prototype is first of its kind developed and evaluated the learning curve on local children in Hyderabad, India.

This paper is organized as follows. Section 2 describes the features of VAA framework. Section 2.1 explains the story selection and recording process. Section 2.2 discuss the dictionary used in this framework. Section 2.3 explains the segmentation algorithm used to automatically align the text and speech. Section 3 discusses the evaluation done using VAA framework.

2. Overview of VAA framework

Fig. 1 shows a screen shot of VAA framework. The interface is designed to give a feel of reading a book and easy to control. The interface consists of controls for – 1) choosing the language of the story, 2) choosing the story for corresponding language selected, 3) play, pause and stop controls for audio, 4) random page selection and 5) navigating to next and previous pages. Additionally, there is audio player embedded to show how much of story is finished reading.

While the goal is to support multiple languages including Indian languages, currently the VAA framework supports English, Telugu, and Kannada. However, the framework can be easily extended to other languages. On choosing the language and the book, the story is loaded into a book-like interface. The user can flip through the pages or can directly go to a required page.

Play button has to be clicked to listen to story. The corresponding lines in the story are highlighted as the story is read. To know a meaning, user has to point the cursor on a particular word. At any point of time, if the user wants to pause or stop, he can click the required controls.

2.1. Selection of audio books

Audio books are available in public domain in multiple languages. Librivox (www.librivox.org) is one such portal where the audio and the corresponding text (from project Gutenberg) could be downloaded. However, a fundamental issue in using such audio books is that they are mostly in English with US/UK accent. Moreover, archaic stories were found to be difficult to comprehend for children.

Hence, stories were selected from a popular Children's magazine called Chandamama. This story book contains short stories and is released in multiple Indian languages including English. A set of parallel stories were identified in English, Telugu and Kannada and were recorded by Indian speakers. Thus each story in English has an associated audio file in Indian accented English. Hence when the student selects the book he wants to read, he can click the play button to hear the story.

2.2. Dictionary

While the story is being read, if the student finds some difficult word, he/she could just role the mouse over the word to get the meaning of that word. The meaning of the words pops up as and when mouse rolls over it and disappears on mouse roll off.

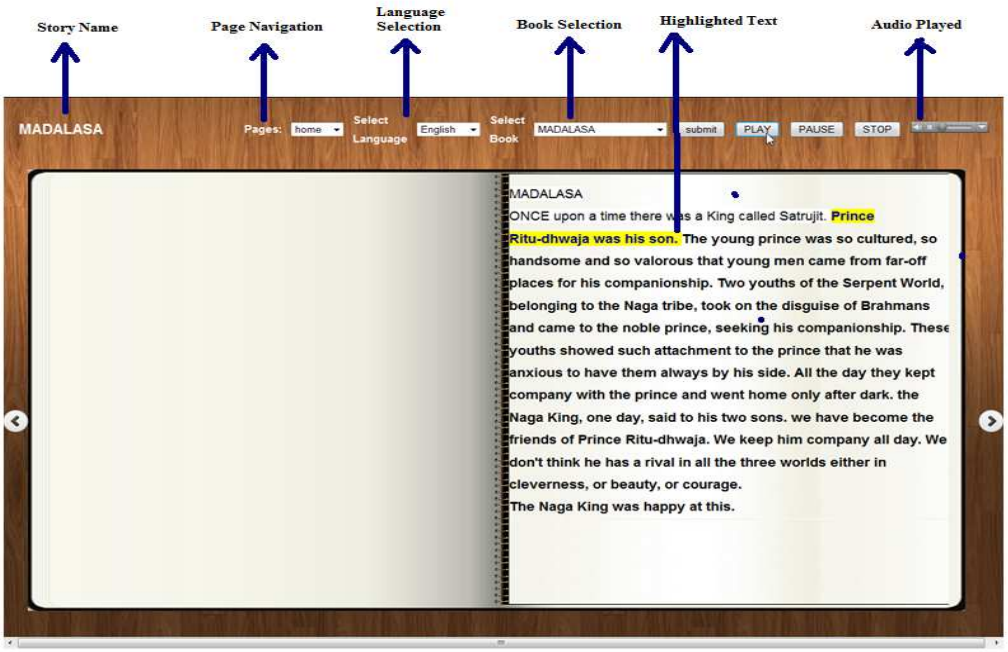

Figure 1: A screen shot of the vocabulary acquisition framework

The dictionary of words is stored in a text file, which is created manually. The words' meanings here are displayed according to the context/story, so as to enable the student to understand both the meaning of the word and the story. The dictionary could be automated by getting the meanings from an online dictionary but we did it manually so that the meanings are according to the context of the story. Also, currently the dictionary is built only for English stories.

2.3. Highlighting words/sentences with voice over

A fundamental issue in highlighting words or sentences at the right time during voice over, is to know beginning and ending time stamps of the words and sentences in the large audio file. This requires automatic alignment of text and audio file. Such alignment is also referred to as segmentation, as the timestamps could be used to segment the audio file. Typically, segmentation can be accomplished by force-aligning an entire audio file with its text using the Viterbi algorithm. However, such solution fails for utterances longer than a few minutes, since memory requirements of the Viterbi algorithm increase with the length of utterances. In [6], an approach is proposed based on modifications to the Viterbi algorithm to process long speech files in parts. In this work, we wish to explore the automatic segmentation of audio books in building a vocabulary acquisition framework. A brief description of this approach is given below.

In the proposed method, the text of large speech file is divided into paragraphs. In an audio book, the text is naturally arranged in paragraphs. Each paragraph consists of one or more sentences, and usually deals with a single thought or topic or quotes a character's continuous words. Let Φ consist of a sequence of K paragraphs $\{u(1), \ldots, u(k), \ldots, u(K)\}$. The words in first paragraph $u(1)$ are force-aligned with first d_u seconds of speech data. As d_u is not known *a priori*, we overestimate its value. Thus a longer speech chunk is force-aligned with the acoustic models[1] corresponding to states in $u(1)$. Let $S = \{1, \ldots, j, \ldots, N\}$ be a state sequence corresponding to the sequence of words in text of the utterance. Let $Y = \{\boldsymbol{y}(1), \boldsymbol{y}(2), \ldots, \boldsymbol{y}(T)\}$ be a sequence of observed feature vectors[2] extracted from the utterance $u(1)$ of T frames. This leads to the case of S emitting $Y' \subset Y$, which could be handled by FA-2 as explained in Section 2.3.1. The result of FA-2 is the correct length of speech chunk corresponding to words in first paragraph $u(1)$. This process is repeated for the remaining paragraphs until the end of text. The sequence of steps involved is as follows:

1. $k = 1, t = 1$.
2. Let $U = [u(k), u(k+1)]$
3. A heuristic estimate of duration of U is defined as $d_u = n_p * d_p$, where n_p is the number of phones in utterance U and d_p denotes the duration of a phone. The value of d_p is chosen as 0.13 seconds such that the estimated value of d_u is *higher* than the actual duration of the utterance U. Let n_f denote the number of frames in d_u seconds and let $F = \{\boldsymbol{y}(t), \boldsymbol{y}(t+1) \ldots, \boldsymbol{y}(t+n_f)\}$ denote the sequence of feature vectors.

[1]The acoustic models used for English to perform segmentation of large audio files are built using about four hours of speech data collected from four CMU ARCTIC speakers (*RMS, BDL, SLT and CLB*).

[2]Speech signal is divided into frames of 10 ms using a frame shift of 5 ms. Each frame of speech data is passed through a set of Mel-frequency filters to obtain 13 cepstral coefficients.

4. Force-align F with the sentence HMM representing U using FA-2 (as explained in Section 2.3.1). As a result of this forced-alignment, the shorter observation sequence $\{\boldsymbol{y}(t), \boldsymbol{y}(t+1)\ldots, \boldsymbol{y}(t+n'_f)\}$ emitted by U is obtained, where $n'_f < n_f$.
5. Given that U is force-aligned with a longer observation sequence, the ending portion of alignment may not be robust – for example, the silence HMM model at the end of U might observe a few observation vectors of next utterance $u(k+2)$, especially if $u(k+2)$ begins with a fricative sound. Hence the observation sequence $\{\boldsymbol{y}(t), \boldsymbol{y}(t+1)\ldots, \boldsymbol{y}(t+n''_f)\}$ corresponding to utterance $u(k)$ alone is considered, where $n''_f < n'_f$.
6. $t = t + n''_f$, $k = k + 1$.
7. Repeat steps 2-6 until $k < K$.
8. In order to obtain phone boundaries for the last utterance $u(K)$ perform forced-alignment of $u(K)$ with $\{\boldsymbol{y}(t), \boldsymbol{y}(t+1)\ldots, \boldsymbol{y}(T)\}$.

2.3.1. Emission of a Shorter Observation Sequence (FA-2)

Let $p(\boldsymbol{y}(t)|x(t) = j)$ denote the emission probability of state j for a feature vector observed at time t and $1 \leq j \leq N$, where N is the total number of states. Let us define $\alpha_t(j) = p(x(t) = j, \boldsymbol{y}(1), \boldsymbol{y}(2), \ldots, \boldsymbol{y}(t))$. This is the joint probability of being in state j at time t, having observed all the acoustic features up to and including time t. This joint probability could be computed frame-by-frame using the recursive equation

$$\alpha_t(j) = \sum_i \alpha_{t-1}(i) a_{i,j} p(\boldsymbol{y}(t)|x(t) = j) \tag{1}$$

where $a_{i,j} = p(x(t) = j|x(t-1) = i)$. Note that Eq. (1) indicates sum of paths, and it transforms to the Viterbi algorithm if the summation is replaced with a $\max$ operation, as shown in Eq. (2).

$$\alpha_t(j) = \max_i \{\alpha_{t-1}(i) a_{i,j}\} p(\boldsymbol{y}(t)|x(t) = j). \tag{2}$$

The values of $a_{i,j}$ and $p(\boldsymbol{y}(t)|x(t) = j)$ are significantly less than 1. For large values of t, $\alpha_t(.)$ tends to zero exponentially, and its computation exceeds the precision range of a machine. Hence $\alpha_t(.)$ values are scaled with term $\frac{1}{\max_i \{\alpha_t(i)\}}$, at every time instant t. This normalization ensures that values of $\alpha_t(.)$ are between 0 and 1 at time t.

Given $\alpha(.)$ values, a backtracking algorithm is used to find the best alignment path. In order to backtrack, an additional variable ϕ is used to store the path as follows:

$$\phi_t(j) = \operatorname*{argmax}_i \{\alpha_{t-1}(i) a_{i,j}\}, \tag{3}$$

where $\phi_t(j)$ denotes a state at time $(t-1)$ which provides an optimal path to reach state j at time t. Given $\phi(.)$ values, a typical backtracking for forced-alignment is as follows:

$$x(T) = N \tag{4}$$
$$x(t) = \phi_{t+1}(x(t+1)),\ t = T-1, T-2, \ldots, 1. \tag{5}$$

When a given state sequence S emits a sequence $Y' \subset Y$, the backtracking part of forced-alignment can be modified as follows. Let $T' < T$ be the length of Y'. To obtain the value of T', the key is to observe the values of $\alpha_t(N)$ for all t. If $1 \leq t << T'$ then $\alpha_{T'}(N) < 1$, and as $t \to T'$ then $\alpha_t(N) \to 1$[3]. This property of state N could be exploited to determine the value of T'. Eq. (6) formally states the property of state N, and could be used to determine the value of T'.

$$\alpha_t(N) \begin{cases} < 1 & 1 \leq t < T' \\ = 1 & t \geq T' \end{cases} \tag{6}$$

Given $T' < T$, the backtracking algorithm is modified as follows.

$$x(T') = N, \tag{7}$$
$$x(t) = \phi_{t+1}(x(t+1)),\ t = T'-1, \ldots, 1. \tag{8}$$

Equation (7) poses the modified constraint that the last state N could be aligned to a feature vector at time $T' < T$, where T' denotes the length of Y' by satisfying Eq. (6). The modified constraint in Eq. (7) allows the backtracking process to pick an observation sequence which is shorter than Y. The forced-alignment algorithm implemented using Eq. (7) and Eq. (8) is henceforth referred to as FA-2.

3. Evaluation

In order to evaluate the effectiveness of this vocabulary acquisition framework, a set of students were selected to use this framework for learning. The children were from grade 2-7 (age 7-13 years). For this study, the subjects were asked to select same stories as others. Also, the evaluation was done only for English stories, as the dictionary was available. The evaluation was performed in two steps referred to as pre-learning and post-learning. During pre-learning, the subject was given a sheet which consists of words from the story. The subject was asked to write down the meanings of the words. Their performance was recorded as score indicating number of words they could correctly answer. This typically shows the knowledge that the subject had before the use of VAA framework. After the subject has finished listening to story and understood the words using VAA framework, the user was given another sheet and was asked to write down the meanings of the same words. This performance was also recorded.

Table 1 shows the performance of subjects in vocabulary acquisition using VAA framework. From Table 1, it can be observed that the average pre-learning score is around 46% while the average post-learning score is around 82%. Fig. 2 shows the performance of the subjects in graphical mode. These results indicate the usefulness of VAA framework for vocabulary acquisition.

4. Conclusions

This work made use of a well known fact in the educational community that children understand and remember the words better when they are explained the meanings based on usage of words, especially in the context of a story. Developing a framework for vocabulary acquisition using audio books requires an alignment of the audio and text. Our contribution has been to automate the alignment process of a large speech file with its

[3]From Eq. (2), it is trivial to observe that a state j achieves an alpha value of 1 at time t, only if it is highly likely to be observed at t. This is dictated by the terms $\max_i \{\alpha_{t-1}(i) a_{i,j}\}$ and $p(\boldsymbol{y}(t)|x(t) = j)$. The alpha value of state N being 1 at time T' implies that the state N is highly likely to be observed at T', and thus the length of observation sequence Y' is T'.

Table 1: *Performance of subjects in vocabulary acquisition using VAA framework.*

S.No	Student Name	Grade/Age	Story Read	Pre-Learning	Post-Learning
1	Mohan	6^{th}/11	The Frog Drum	5/10	10/10
2	Aarthi	4^{th}/9	Swindlers	4/10	9/10
3	Venu	6^{th}/11	Strange Painting	6/10	10/10
			Good Eater	7/9	9/9
			Swindlers	6/10	9/10
4	Kruthi	4^{th}/9	Madalasa	4/9	7/9
			Good Eater	6/9	6/9
5	Rithika	5^{th}/10	Swindlers	4/10	7/10
			The Frog Drum	3/10	8/10
6	Ananya	3^{rd}/8	Strange Painting	3/11	7/11
7	Anjana	8^{th}/13	Loosing Friends	6/12	9/12
8	Aakruthi	5^{th}/10	Loosing Friends	2/12	10/12
9	Rahul	7^{th}/12	Sindabad The Sailor	8/10	10/10
10	Arnav	2^{nd}/7	The Strange Painting	2/10	5/10
			Average Score (in %)	45.6%	82.9%

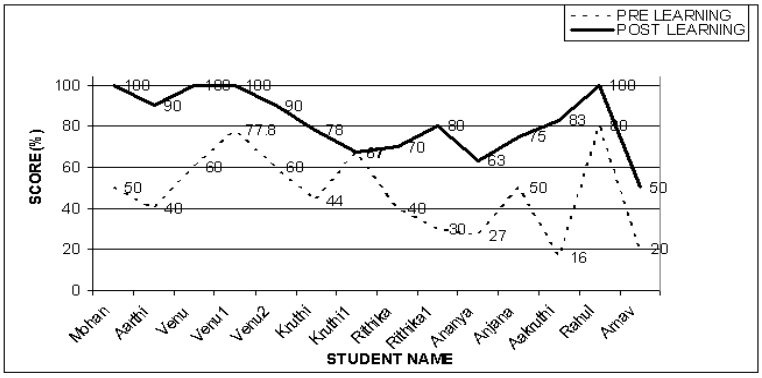

Figure 2: Performance of subjects using vocabulary acquisition framework

corresponding transcription, and to build a framework for vocabulary acquisition.

Using this prototype framework, we have tested the vocabulary acquisition of children in the locality of Hyderabad, Andhra Pradesh, India. We have shown that use of short moral stories helps in improving the vocabulary of children. Using this framework, the learners will get to know the meanings of the words as they run through the story. As far as our knowledge, such a project is first of its kind developed and tested in Andhra Pradesh, India. It should be noted the current validation of this framework has not been strong, i.e., there is no control group and statistics on the learning. We plan to conduct a more rigorous validation of this framework using a control group and a large number of audio books.

5. Acknowledgements

We would like to thank Prof. Raj Reddy, Carnegie Mellon University for initial discussions on language learning using audio books.

6. References

[1] M. Heilman, L. Zhao, J. Pino, and M. Eskenazi, "Retrieval of reading materials for vocabulary and reading practice," in 3^{rd} *Workshop on Innovative Use of NLP for Building Educational Applications*, Columbus, Ohio, USA, 2008.

[2] L. Liu, J. Mostow, and G. Aist, "Automated generation of example contexts for helping children learn vocabulary," in *Second ISCA Workshop on Speech and Language Technology in Education*, University of Birmingham, UK, 2009.

[3] I. McGraw, B. Yoshimoto, and S. Seneff, "Speeech-enabled card games for incidental vocabulary acquisition in a foreign language," *Speech Communication*, vol. 51, no. 10, pp. 1006–1023, 2009.

[4] B. Laufer and J. Hulstijn, "Incidental vocabulary acquisition in a second language: The construct of task-induced involvement," *Applied Linguistics*, vol. 22, pp. 1–26, 2001.

[5] W. B. Elley, "Vocabulary acquisition from listening to stories," *Reading Research Quarterly*, vol. 24, pp. 174–87, 1989.

[6] K. Prahallad and A. W. Black, "Segmentation of monologues in audio books for building synthetic voices," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 5, pp. 1444–1449, 2011.