

Automatic Audio Indexing and Audio Playback Speed Control as Tools for Language Learning

David Rossiter, Gibson Lam, and Brian Mak

Department of Computer Science and Engineering,
Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
{rossiter, gibson, mak}@cse.ust.hk

Abstract. The Gong system has been developed for web based communication. It supports synchronous and asynchronous audio communication and can be embedded in other learning management systems. This paper discusses two novel features which are targeted at language learners using the system. The first is the ability to automatically index an audio recording. After the indexing has taken place the user is able to select one or several words and hear just those words spoken in isolation. The second is the ability to vary the playback speed of any recorded message. The technical details of their implementation as well as pedagogical use of these features are discussed.

Keywords: Audio Indexing, e-Learning Tools, Speech Recognition.

1 Introduction

The Gong system has been created as a tool for web based communication and learning. The system supports both synchronous and asynchronous voice and text communication. The input and display of Unicode characters is supported, as well as specially developed features for the input and display of the Cantonese and Putonghua dialects of Chinese. The entire system can operate in a number of different languages, and can be embedded inside other learning management systems. An API has been developed so that the features of Gong can be accessed and controlled in the web page environment by languages such as JavaScript, VBScript and Flash. A module has been released so that Gong can be used as an integrated component in the popular Moodle learning management system. See [1] for a description of Chinese language features as well as a general overview of the Gong system. Our web site [2] may be accessed for more information about the system including what it can do and how it can be used, and for downloading Gong.

An example of the system being used as a voice board is shown in figure 1. The display shows messages in the voice board. The currently selected message is shown under the message list. A list of currently logged on users is also shown, facilitating real-time audio and text chat.

In addition, special features have been developed which are intended primarily for language learners. The main focus of this paper is on two such features: the automatic indexing of spoken audio, and audio playback speed control.

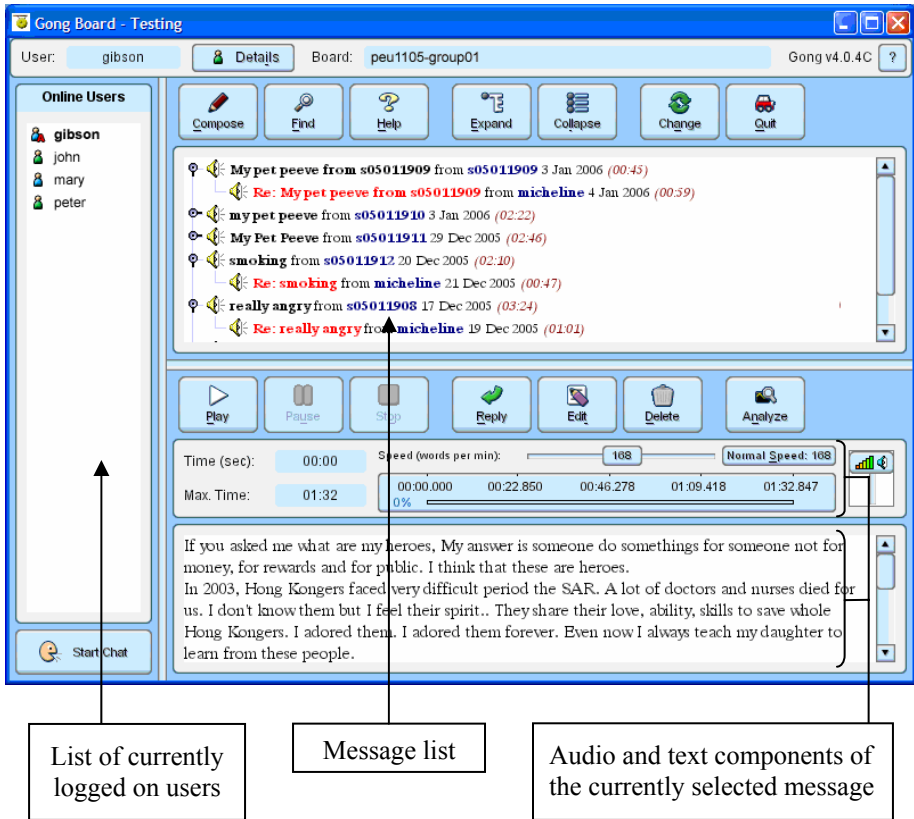


Fig. 1. A typical Gong voice board display

2 The Gong System

The Gong system is built using Java. The system requires the Tomcat server for server side operation. For client side deployment Gong is available as an applet as well as an application. Functionally, both are equivalent. For applet operation the Java run-time environment must be previously installed in the client system. For operation as an application there is no such requirement, as a preferred version of Java is included within the installation. When used as a component in other learning management systems Gong is used as an applet.

3 Indexed Audio

3.1 Overview

In the most common use of the Gong system, a text and/or audio message is created by a user and is posted to a voice board where others can read and give feedback on it.

We have enhanced our system with the ability to automatically index any message. That is, when the text is a transcription of the spoken words of the same message, then we can automatically match the words to their spoken counterparts. The indexing can then be used for a number of pedagogical functions.

From the user's point of view the generation of the word indexing is simple. The user selects a button in order to create a new message. He/she then records the speech component of the message. The corresponding text is entered. Finally, the user selects 'Index'. After a short delay the automatic audio indexing process is complete. The message can then be posted so that any users can use the pedagogical functions that are built upon the results of the audio indexing operation.

3.2 Pedagogical Usage

The indexing results in a sequence of start and end tags indicating the start and end of the corresponding spoken word. Figure 2 illustrates the start of word tags for the four words 'she eclipses and predominates' overlaid on a spectrogram of the speech signal.

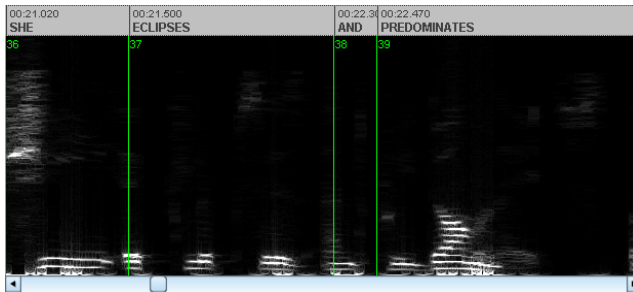


Fig. 2. The display of starting word markers on top of a spectrogram of the speech

From a pedagogical point of view there are a number of ways in which the indexed audio may be used.

- During playback of all or part of the audio the words which are being played at that moment in time are highlighted while they are being played. This greatly assists learners in developing sight reading skills, as there is a clear visual indication of the progress of the speech.
- The user may double click on a single word in order to hear it in isolation, or the user may select any number of words and press play in order to hear those words. This is illustrated in figure 3. In this way the user can concentrate on phrases which require repeated attention, without being required to play the entire speech again.
- The user may search for any word or phrase within a message. All instances of the search term are then shown highlighted, as indicated in figure 4. The words can therefore be easily played and contrasted with each, in order to reveal contextual differences in pronunciation.

The phrase 'In his eyes' is selected for selective playback

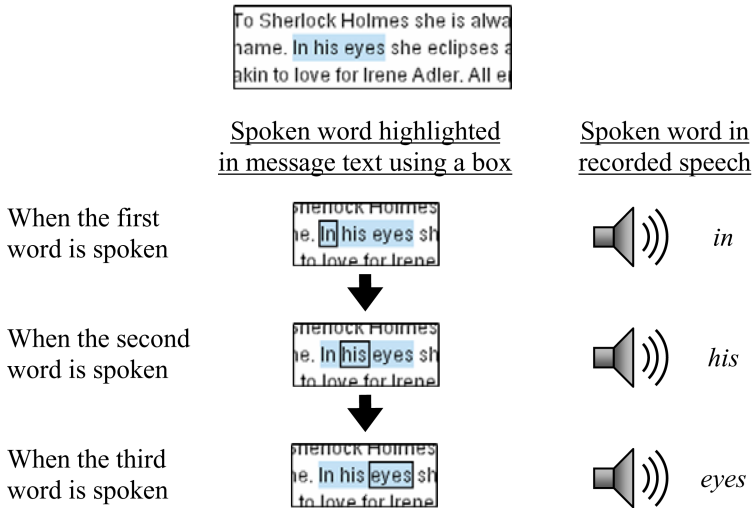


Fig. 3. An example of playing back three selected words, after audio indexing has taken place

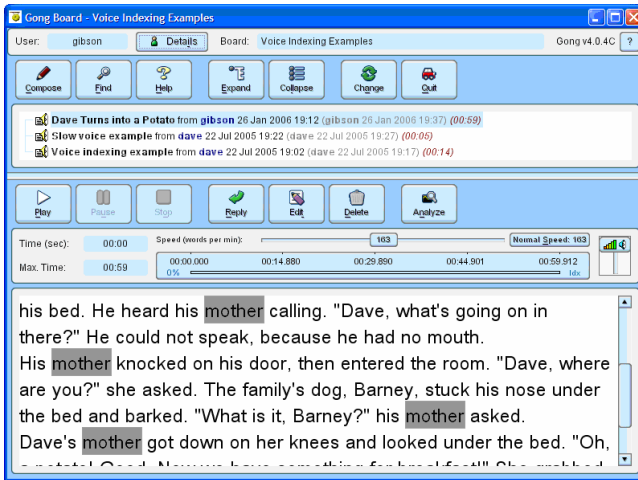


Fig. 4. An example of message text search. In this case four instances of the search term 'mother' have been highlighted by the system, and can be individually selected for audio playback.

3.3 Methodology

The audio indexing is performed by an automatic speech recognition (ASR) system. When Gong is being used as an applet the entire recognition system is downloaded into the client machine when the text indexing operation begins. This dynamic download ensures faster download of the applet, as audio indexing may not be used

every time. For the application version of the Gong client side code the ASR is included in the installation.

A simple overview of the automatic speech recognition (ASR) system is shown in figure 5.

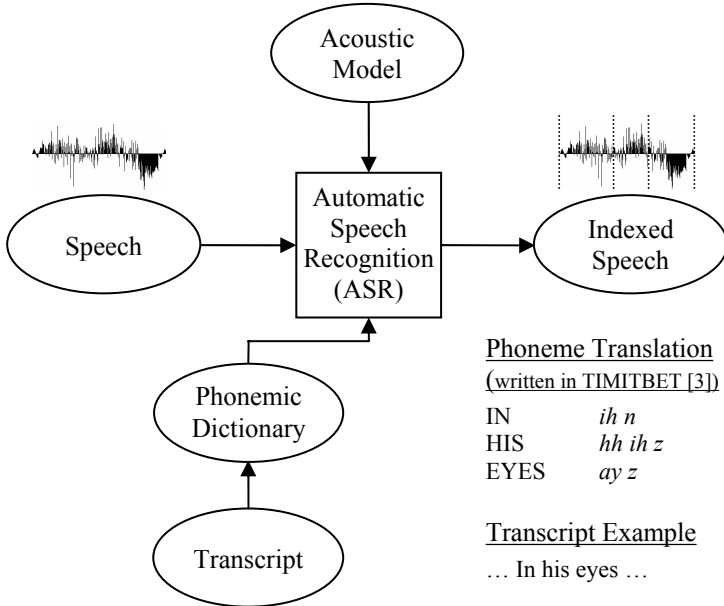


Fig. 5. An overview of the automatic speech recognition system

The ASR system consists of

- A dictionary file. This contains the phonemic transcriptions of more than 100,000 common English words.
- A set of acoustic models, one for each English phoneme. We use a set of 39 English phonemes to label all English words in the dictionary. They are implemented as hidden Markov models (HMMs) [4]. Hidden Markov modeling is the most commonly used technology in today's state-of-the-art ASR systems. An HMM is basically a stochastic finite-state automaton, and the probability that an acoustic observation is generated by an HMM state is governed by a probability distribution. In our system each HMM has 3 states arranged from left to right with no skipping or reversing arcs, as shown in figure 6. The 3 states are used to capture the acoustics at the onset, middle, and ending of a phoneme during its realisation.
- A phoneme recogniser. This can identify the phonemic contents in an utterance. The recogniser in our system employs statistical pattern recognition techniques.

The recogniser in Gong works as follows: the corresponding phonemic transcription is first created by looking up each word within the orthographic transcription (that is, the textual component of the Gong message) with our electronic

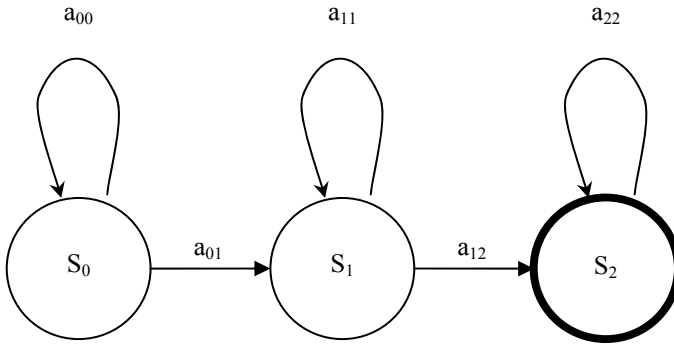


Fig. 6. The three states of a HMM used in our model

<p>Voice data display selection Controls for selecting spectrogram or waveform display.</p>	<p>Acoustic model Selection control for the acoustic model to be used by the speech recogniser.</p>
	<p>Spectrogram bin-size Selection control for the FFT bin-size. Available values range from 16 to 512.</p>
	<p>Word indexes & timing The bar shows the time for each word. The start of word index is shown as a vertical line super-imposed on the spectrogram.</p>
	<p>Voice data display The audio recording is displayed here. It can be shown as a spectrogram or as a waveform according to the user's choice.</p>
<p>Text content The text component of the message, which matches the spoken audio component.</p>	<p>Speed adjustment slider The audio playback speed can be adjusted, with a 'words-per-minute' measure automatically updated.</p>

Fig. 7. An overview of the user interface of the automatic speech recognition component of the system

dictionary. Based on the phonemic transcription and with the use of the acoustic models, the phoneme recogniser will find the optimal time alignment between the phonemic transcription and the audio recording. In other words, the recogniser tries to find out the best beginning time marker and ending time marker for each word in the given utterance. The aligning procedure involves warping the acoustic signal dynamically along the time domain to fit the corresponding phoneme models until we

get the most likely alignment - in the statistical sense - using a dynamic programming algorithm called the Viterbi search [5].

The Gong system supports any number of models, so it would be possible to support audio indexing of different languages. At the time of writing only one model and phoneme dictionary is used, for the English language.

Figure 7 shows the complete display in the Gong system for interfacing with the ASR sub-system. Controls for the selection of the acoustic model and spectrogram display are shown near the top, followed by a spectrogram display of the speech, and then the textual equivalent. The elements near the top of the GUI which are related to models and spectrograms are kept hidden by default in order not to scare non-technical users of the system.

3.4 Words-Per-Minute Metric

After a message has been indexed a words-per-minute value for the message is shown. This is obtained by simply dividing the total number of words in the message by the duration of the spoken message (in seconds), and then multiplying by 60. This value provides an absolute speed measure which by itself can be used to provide an indication of the speed of the speech. For a language learner the difficulty of comprehension of the message is proportional to the number of words spoken per minute. Reference values can be used to help the learner rate his/ her level of understanding. For example, quality presentations are typically given at 120-170 words per minute; a normal conversation by native English speakers is approximately 200 words per minute; auctioneers use around 250 words per minute.

4 Audio Playback Speed Control

4.1 Overview

The words-per-minute value is combined with another feature of the Gong client, which is the ability to speed up or slow down playback of the voice recording without

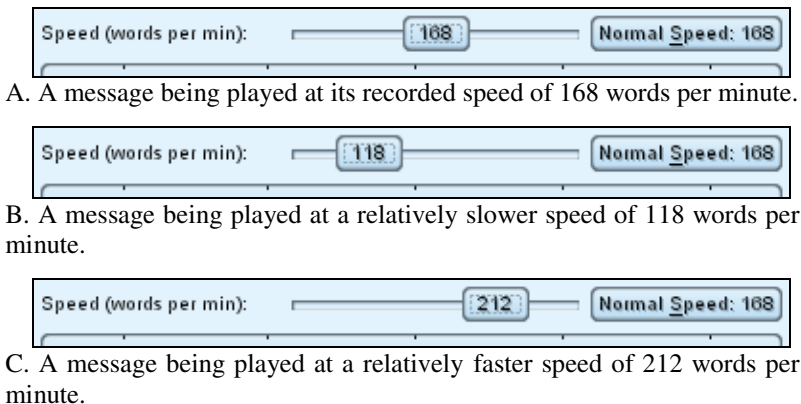
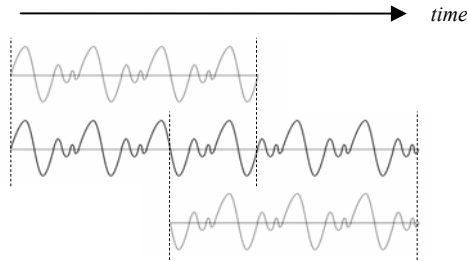
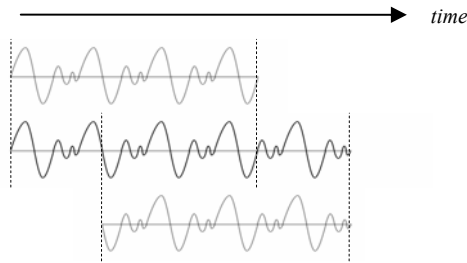


Fig. 8. An example of the words-per-minute value used as a measure for audio playback speed control

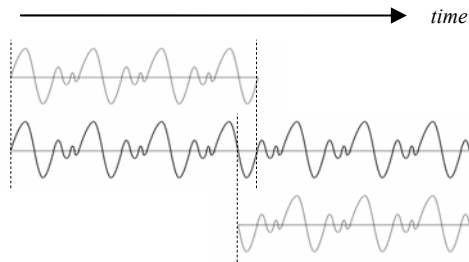
otherwise changing the sound. Usage of this feature is illustrated in figure 8. As the speed control slider is moved the numerical display of the words-per-minute measure is automatically altered to accurately reflect the new playback speed in terms of words-per-minute.



1. The audio recording (the middle trace) is divided into two overlapping sections (the top and the bottom one).



2a. Speeding Up: The second section is moved closer to the first one and therefore the resulting audio (the middle trace) becomes shorter.



2b. Slowing Down: The overlapping region is shortened and therefore the resulting audio (the middle trace) becomes longer.

Fig. 9. An illustration of how the speed of an audio recording is altered without changing the pitch

The playback speed control is a real-time operation meaning that users can interactively select a new speed and immediately hear the result.

4.2 Pedagogical Usage

An example of the use of speeding up the audio would be when a teacher is reviewing a language learning student's recorded voice, and the student is speaking very slowly. An example of the use of slowing down the audio would be when the student is trying to understand the spoken English of a native speaker speaking at a natural speed. The automatically adjusted words-per-minute measure provides a value by which the student can judge his/her progress over time, as increasing words-per-minute speeds become more comprehensible to the student.

4.3 Methodology

The algorithm used to change the speed of the recorded speech without seeming to adversely change the speech is based on the Synchronized Overlap Add (SOLA) method for speeding up and slowing down digital audio recordings [6].

Figure 9 provides an illustration of how this method is used to change the speed of the speech recording. The digital audio is first divided into a series of overlapping sections. These sections are then re-combined by adjusting the size of the overlapping region. The method tries to find the best overlapping position by maximizing a cross-correlation function so that there is minimal distortion in the resulting signal. The overlapping region of different sections is then combined using a cross-fade method.

To speed up digital audio, the size of the overlapping region is increased and therefore the length of the audio result becomes shorter. Relatively shorter length of the audio results in relatively faster audio playback. In contrast, to slow down an audio recording, the overlapping regions among these overlapping sections are shortened so as to increase the overall length of the audio recording. Because of the relatively longer length of the audio it produces a relatively slower playback.

5 Usage and Availability

The features described in this paper have been used as part of Gong by students and teachers at several education institutions including the author's home institute, distance learning institutions and secondary schools. At this stage we have not done a formal study of the advantages of the features, but informal observation shows that they clearly assist language learners in learning English.

6 Conclusions

The Gong system is a web based communication tool with support for learning. We have introduced two features which are especially useful for language learners. The first one is audio indexing. Given a voice recording and its transcript it allows learners to selectively playback a single word or a phrase by simply selecting the words. The second feature is audio playback speed control. Learners can playback a voice

recording faster or slower measured by a words-per-minute metric. We have explained the methodology of these two features as well as various pedagogical usages depending on different situations.

Acknowledgments. The Gong project has been supported by several Continuous Learning and Improvement (CLI) grants, as well as a grant from the Vice President for Academic Affairs office (VPAAO) of the Hong Kong University of Science and Technology. CLI grants are funded by a Hong Kong Teaching and Development grant obtained by the Centre for Enhanced Learning and Teaching (CELT) division of the University.

References

1. Rossiter, D., Lam, G., Cheng, V.: The Gong System: Web-based Learning for Multiple Languages, with Special Support for the Yale Representation of Cantonese, *The 4th International Conference on Web-based Learning*, 31st July - 3rd August 2005, Hong Kong SAR, China, *Lecture Notes in Computer Science* vol. 3583, Springer-Verlag, (2005)209-220.
2. The Gong Project web site, <http://gong.ust.hk>
3. Zue, V., Seneff, S., Glass, J.: Speech database development at MIT: TIMIT and beyond, *Speech Communication*, Vol. 9, No. 6, August (1990) 351-356.
4. Rabiner, L.A.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”, *Proceedings of the IEEE*, Vol. 77, No. 2, February (1989) 257-285.
5. Forney, G. D.: The Viterbi Algorithm”, *Proceedings of the IEEE*, Vol. 61, No. 3, March (1973) 268-278.
6. Roucos, S., Wilgus, A.M.: High Quality Time Scale Modification for Speech, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, March (1985) 493-496.