THAI LANGUAGE AUDIO RESOURCE CENTER:
THAI SPEECH DATABASE AND APPLICATION
IN WEB-BASED LANGUAGE TEACHING

Yupaphann Hoonchamlong
University of Hawaii at Manoa
<yupapha@hawaii.edu>

Sathaporn Koraksawet
Thammasat University
<stp@tu.ac.th>

Rugchanok Janevarakul
Thammasat University
<rugchanokj@hotmail.com>

0 Introduction
Recently, the importance of corpora or databases of language in research in linguistics, lexicography and natural language processing has gained increased recognition. The growth of the high speed internet enables the dissemination of and access to these text and multimedia language resources with ease and speed, making resource sharing among researchers from every corner of the world a possibility.

The development of audio resources of Thai language (especially speech) for research and development with dissemination via the internet is still in an early stage of evolution. This paper reports on Thai Language Audio Resource Center, one such Thai language resource now available via the web.(Hoonchamlong et al. 2002a, 2002b)

Thai Language Audio Resource Center (ThaiARC) project provides audio information (in the form of digitized audio files) on Thai language (especially speech) for academic research, disseminated via the Web as part of ThaiSARN (Thai Social/Scientific Academic Research Network) at http://thaiarc.tu.ac.th. ThaiARC also serves as an archive for these collections as reference and as a shared resource on Thai language for researchers.

The project pioneers the production and collection of various types of audio information on Thailand and the Thai language, such as royal speeches, academic lectures, oral literature, etc, for dissemination on the web. It also explores the application of speech data in web-based language teaching.

The audio files at ThaiARC are available in Wave (.wav) RealAudio (.ra) and mp3 formats, formats that are readily accessible by most computers. Besides being a repository for the collections of audio information on Thai Language, ThaiARC also provides electronic verbatim text/transcript and annotation accompanying each audio program, also distributed electronically through the web.

ThaiARC was funded by NECTEC (National Electronic and Computer Technology Center), NSTDA (National Science and Technology Development Agency), Ministry of Science, Technology and Environment of Thailand during the year 1997-2002. The project
was carried out in two phases at Thammasat University which also hosts the project website.

Phase 1 of the project was conducted during 1997-1999 as a pilot phase to establish the audio resource center by studying appropriate technology to be used in each stage of database development, namely, data formatting, archiving, data accessing and retrieval. Samples of three groups of Thai audio information, namely, speech styles, regional folktales and poetry are made available. In addition, a Thai Language Page was also established especially to provide English information on the Thai language to foreigners.

Phase 2 of the project was conducted during 2000-2002. The objectives of Thai Language Audio Resource Center (THAIARC) project Phase 2 are

1) to establish a resource for Thai language audio information for research, by systematically collecting Thai audio data for linguistic research, starting from the tones of the four Thai dialects, which are the distinct characteristics of Thai language. The tonal word sets are collected according to linguistics research design for tones.

2) to investigate the forms and techniques most appropriate for web-based language instruction, and to apply the collected Thai audio information in developing a model web-based lesson for teaching Thai listening comprehension skills for foreigners.

At present, the following four groups of Thai speech data are available as an online voice sample library on the following topics, along with verbatim transcripts and annotations:

1) Thai Regional Dialects: samples of word sets demonstrating tonal variations among the 4 major Thai dialects (Northern, Northeastern, Southern, Central) and Standard Thai, with audio file search tool for easy access, located at: http://thaiarc.tu.ac.th/host/thaiarc/dialect

2) Thai Regional Folktales: samples of Thai Folktales from the 4 major regions (Northern, Northeastern, Southern, Central), each with transcript and standard Thai translation.

3) Thai Poetry: samples of readings of various poetic styles and versifications. Explanations are available in both Thai and English.

4) Speech Styles: samples of the King’s speeches for various occasions, for example, his Golden Jubilee Speech; also here are samples of various types of news broadcasts.

THAIARC also features the “Thai Language Page” which aims to provide overview information about Thai language in various aspects for foreigners. This information includes the history of the Thai language, Thai alphabets and basic Thai phrases with audio files for tourists or those interested in Thai.

In addition, a sample application of THAIARC audio and transcript data as “Web-based Instruction of Thai Listening Skills” is also demonstrated.
Data Collection, Archiving and Dissemination

1 Thai Regional Dialects Word Sets

Thai is a tonal language belonging to the Tai language family, which includes languages spoken in Assam, northern Burma, all of Thailand including the peninsula, Laos, Northern Vietnam and the Chinese provinces of Yunnan, Guizhou (Kweichow) and Guangxi(Kwangsi).

In Thai and Tai dialectology, tone systems and variations in tone systems are important identifying features of the various Thai/Tai dialects. Based on a method developed in the discipline of comparative and historical Thai linguistics, William Gedney (1972)’s “Proto Tai Tone Matrix” or commonly known as “Gedney’s Tone Box” has been widely used as a tool for collecting word lists for determining the tone system of a Thai/Tai dialect (see Figure 2.). The tone box is based on the development of various Tai tone systems by tone splits and tone mergers from the reconstructed three tones (Tone A, B and C) of Proto-Tai. The tone splits and mergers were influenced by certain features of initial consonants of Proto-Tai such as voicing and aspiration, and also the types of syllables and vowel length. The word lists collected are words that are known to be Tai in origin, i.e. words describing daily life and environment in villages shared by related Tai dialects.
### Figure 2: Proto-Tai tone matrix

The following diagram (Figure 3), adapted from Tingsabadh 2001, show how Standard Thai consonant classes and tone markers correspond with Proto-Tai initial consonant classes and tones.

<table>
<thead>
<tr>
<th>Proto Tai Initial Consonants at time of split</th>
<th>Standard Thai Consonants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1. Voiceless Friction</td>
<td>&quot;HIGH&quot;</td>
</tr>
<tr>
<td>*s *hm *ph</td>
<td>ข ฃ ฅ ฆ ช ฌ ญ ฑ ฒ ฑฆะ</td>
</tr>
<tr>
<td></td>
<td>kh ch th ph f s h m n ย ร ล ว</td>
</tr>
<tr>
<td>Class 2. Voiceless unaspirated stops</td>
<td>&quot;MID&quot;</td>
</tr>
<tr>
<td>*p *t</td>
<td>ก จ ต ฏ</td>
</tr>
<tr>
<td></td>
<td>k e t p</td>
</tr>
<tr>
<td>Class 3. Glottal</td>
<td>&quot;LOW&quot;</td>
</tr>
<tr>
<td>*ʔ *ʔb</td>
<td>บ ด ฉ</td>
</tr>
<tr>
<td></td>
<td>b d ?</td>
</tr>
<tr>
<td>Class 4. Voiced</td>
<td></td>
</tr>
<tr>
<td>*b *m</td>
<td>ด ต ดฆ ตป</td>
</tr>
<tr>
<td></td>
<td>kh ch th ph f s h m n ย ร ล ว</td>
</tr>
</tbody>
</table>

### Figure 3: Proto Tai and Standard Thai consonant classes and tones

The following (Figure 4.) show the word lists, based on the Tone Box, that we use for collecting dialectal word samples.