DONLabel: An Automatic Labeling Tool for Indian Languages

P. G. Deivapalan, Mukund Jha, Rakesh Guttikonda, Hema. A. Murthy
Department of Computer Science and Engg.
Indian Institute of Technology Madras
Chennai, India 600036
Email: {deivapalan,mukundj,rakesh,hema}@lantana.tenet.res.in

***Abstract*—Although, continuous speech recognizer (CSR) and speech synthesis systems have been developed for a number of languages in the world, such systems are not available for Indian languages. The primary reason for this is a lack of annotated databases for the speech. In this paper, we describe DONLabel, an automatic labeling tool for Indian languages. This tool enables segmentation of a speech signal at syllable boundaries which is then labeled with the corresponding text. As segmentation might not be always accurate. The tool also has facilities for correcting the boundaries.**

I. Introduction

Adapting an existing speech synthesis or speech recognition systems to a new language or to a new task require building the systems on a large labeled database. The bottleneck is not in collecting database, but labeling the database. Typically, labeling involves defining a speech segment and assigning a suitable label to it. As most Indian languages are syllable-centric in nature, syllable could be an appropriate segment to label. Manual labeling is not only laborious, but also time-consuming. Also, it might introduce errors in the definition of a segment. The correctness of labeling depends upon the skill and experience of the person involved. It is necessary to automate this process, so that the time taken to build a new system is significantly reduced. At the same time, allowing human experts to correct the mistakes (wrongly marked segments) would improve the accuracy of labeling. Existing automatic tools [1], [2] provide labels only at the phonemic-level using trained acoustic models. But systems [3], [4] built for Indian languages require labels at the syllable-level. Hence, it is important to build an automatic labeling tool that enables the database to be labeled in terms of syllables.

In a CSR for Indian languages that was built at IIT Madras [4], it was shown that speech signal annotated at syllable-level perform well for Indian languages. In [4], [5], a group deley based segmentation algorithm was developed to segment speech signal at the syllable-level. In this paper, we describe a GUI-based tool, which is built around this setup, that performs automatic labeling. The main contribution is that, this tool does not require any trained acoustic models to perform segmentation of speech signal. To our knowledge, this is the first attempts to develop such a tool for Indian languages.

Labeling large amounts of speech corpora requires a suitable software environment. The tool described in this paper is called DONLabel. It provides the following facilities:

1) automatic segmentation and labeling
2) does not require any trained acoustic models for speech segmentation
3) allows segment boundaries to be added or removed or changed
4) currently displays the labeled text in two Indian languages either Tamil or Hindi
5) plays a segment of speech
6) zooms a selected speech segment
7) facilities to modify parameters for the segmentation.

It is based on a simple architecture which contains input unit, processing unit and output unit. A web-based interface has also been provided for this tool. In this case, the processing unit runs in the server while the input and output units are handled through a web browser (client). Nevertheless, all the units can run on the same computer for a stand-alone usage.

The rest of this paper is organized as follows: Section II explains the background and system requirements of DONLabel. In Section II-B, existing software are discussed. Section III explains the tool design. In Section IV, DONLabel is evaluated on the database for Indian languages. Section V concludes this paper.

II. Background

We have defined our needs concerning a tool for labeling of speech corpora. We have noticed that none of the existing programs were really adequate, so we decided to develop a new one.

A. Requirements

The requirements for the transcription tool were the following:

- handling long audio signals (approx 60 sec);
- displays text in different Indian languages;
- able to use the tool without installing it;
- usable by a non-specialist.

Concerning the last point and the user interface, we thought it would be particularly important to provide an interactive access with zoom options.

B. Existing labeling software

Earlier efforts [1], [2], [6], [7], have produced similar labeling tools. The currently available tools can be broadly

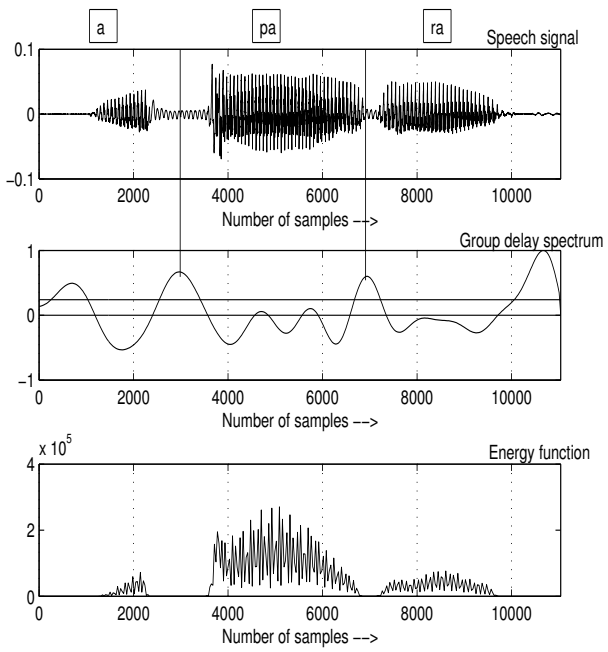

Fig. 1. Group delay segmentation of utterance *ap-pa-ra*.

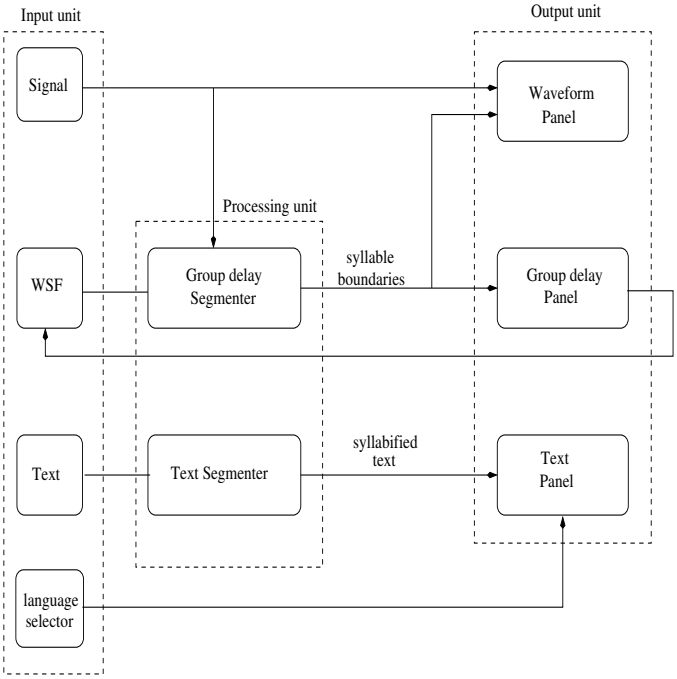

Fig. 2. Functional block diagram of labeling tool

classified into two categories, namely GUI tools and non-GUI tools. In GUI tools, the labeling is performed manually ie., user can mark the segment boundaries. It allows to change the boundaries. for example, a boundary can be added or deleted or moved. Emulabel [6] and wavesurfer [7] come under this category. In non-GUI tools, the labeling is performed automatically. The acoustic models are trained from the data and are used for segmentation. Once labeled, the boundaries are fixed. Festvox [1] and SPHINX [2] come under this category.

C. Group delay based segmentation

The negative derivative of the Fourier transform phase is defined as "group delay". It has been shown in [8], [9] that, the short-term energy function of the speech signal can be processed in the group delay domain to produce syllable segments. The resolution of the segmentation algorithm is dependent on a parameter called window scale factor (WSF) [10]. Large values of WSF result in merged syllable boundaries, while small values of WSF result in syllables that are split. The group delay based segmentation is illustrated in Fig. 1.

III. Tool design

The DONLabel interface is developed in Java [11], where as the segmentation engines are developed in C. There are two versions, namely the stand-alone version and the web based version. The stand-alone version can be installed in the Linux platform. The web-based version of this tool is available at: http://www.lantana.tenet.res.in/apache2-default/main.php.

A. Architecture

The DONLabel consists of three units namely (i) input unit, (ii) processing unit, and (iii) output unit. The functional block diagram is shown in Fig. 2. Each unit contains sub-units which are explained below.

The input unit takes four parameters namely (i) waveform, the given speech signal which has to be labeled, (ii) transcription, the text corresponding to the given speech signal, (iii) language, a Indian language in which the label has to be displayed, and (iv) parameters for labeling engine.

This unit consists of two sub-units namely (i) group delay segmenter and (ii) text segmenter. Each of these sub-units are explained below.

The *Group delay segmenter* performs the speech segmentation. The given speech signal is segmented using group delay based segmentation algorithm [12]. The syllable boundary information is passed to the output unit for further processing. The WSF plays an important role in obtaining proper and better segment boundaries.

The *Text segmenter* is used to perform syllabification. The transcription corresponding to the spoken utterance is syllabified using rule based approach [10]. The syllabified text and the chosen Indian language is passed to the output unit.

The output unit consists of three panels namely (i) text panel, (ii) waveform panel, and (iii) group delay panel. Each of these panels are explained below.

The text panel displays the syllabified text in Indian language. The group delay panel displays group delay function of the speech signal. The wave panel displays the given speech signal along with the syllable boundaries present in it (see Fig. 3).

B. Audio playback server

Audio playback is also provided by a server controlled through sockets. This allows:

- remote execution of the tool with playback on the display terminal;
- concurrent access of several clients to the sound driver;

C. User interface

1) sound viewer: The temporal shape of sound signal is displayed in a window. Real-time scale modification from a global overview of the signal to a sample-by-sample view is possible. Segments can be added or deleted or moved with simple mouse movements in the signal window.

The text panel displays the syllabified text in Indian language. The group delay panel displays the group delay function of the speech signal. The wave panel displays the given speech signal along with the syllable boundaries present in it (see Fig. 3).

2) zoom in and zoom out: The zoom button (see Fig. 3) is used to zoom a particular segment of waveform. While zooming , the alignment of the group delay panels and the text panels are maintained. The tool provides three kinds of zoom options namely (i) zoom in, (ii) zoom out, and (iii) zoom to fit the screen.

3) Add or remove or change boundary lines: Addition of boundary lines are performed by clicking the left mouse button. Removal of boundary lines are performed by clicking the right mouse button. DONLabel allows user to move the boundary lines. By left clicking on the *diamond shape* (see Fig. 3) and dragging the mouse to a desired location, one can actually move the boundary lines.

4) Play waveform: DONLabel allow the user to play the entire waveform by pressing the leftmost play button (see Fig. 3). It also supports play of a particular segment selected by the user. Selection of a segment is performed by dragging the mouse with the left button pressed and then pressing the second play button from the left (see Fig. 3).

5) Save as lab file: The information such as duration of each segment along with the label of each segments can be stored by pressing the save button (see Fig. 3). The lab file format (see Fig. 4.) is similar to the format used in other tools except that labels are displayed in Indian languages. The first three lines forms the header information. Remaining each line contains the duration of the syllable, for eg., the start and end duration of "nak" is 0.15375 sec and 0.255 sec respectively (as shown in Fig. 4).

IV. Evaluation of DONLabel

A single bulletin from Database for Indian language (DBIL) [13] is selected for the task of labeling. It contains 120 sentences and 1316 syllabified segments. DONLabel is used for labeling. It consists of 1288 group delay segments and 1316 text segments. Inorder to make the evaluation interesting, the task of labeling the single bulletin is given to a novice user (see acknowledgment).

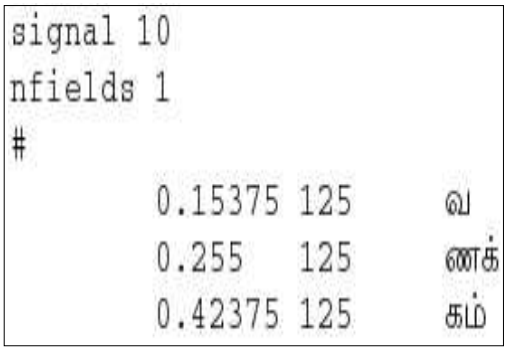

Fig. 4. A sample lab file represented using standard format.

TABLE I
Performance of the labeling tool on DDNEWS bulletin.

	syllable segments
Number of correct	1251
Number of insertion errors	37
Number of deletion errors	65
Tot. Number of GD segments	1288
Tot. Number of Txt segments	1316

The result is shown in Table I. The accuracy of the segmentation is found to be 97.2%. It should be noted that, we obtained this performance after tuning the WSF with different values. Also notice that there might be some boundary misaligned.

V. Conclusion

In this paper, we have attempted to provide a solution to label speech databases for Indian languages. A GUI-based automatic labeling tool for Indian languages has been built. A web-based interface is provided to the end-user so that without installing this tool, they can perform labeling task. It has been evaluated on a single bulletin from DBIL. The results show an accuracy of 97.2% with very small insertion and deletion errors.

Acknowledgment

The authors would like to thank Elumazhai of DONlab, IIT Madras for helping in the experiments.

References

[1] Festvox [Online]. Available: http://festvox.org/
[2] The CMU Sphinx Group Open Source Speech Recognition Engines [Online]. Available: http://cmusphinx.sourceforge.net/html/cmusphinx.php
[3] Samuel Thomas, M. Nageshwara Rao, Hema. A. Murthy, C. S. Ramalingam, "Natural Sounding TTS based on Syllable-like Units," in *Proc. of 14th European Signal Processing Conference*, Florence, Italy, Sep. 2006.
[4] Lakshmi, "A Syllable based Continuous Speech Recognizer for Tamil", M.S. thesis, Indian Institute of Technology Madras, Chennai, July 2007.
[5] P. G. Deivapalan and Hema. A. Murthy, "A syllable-based IWR for Tamil handling OOV words", (pre-print).
[6] The EMU speech database system [Online]. Available: http://emu.sourceforge.net/
[7] Wavesurfer [Online]. Available: http://www.speech.kth.se/wavesurfer/

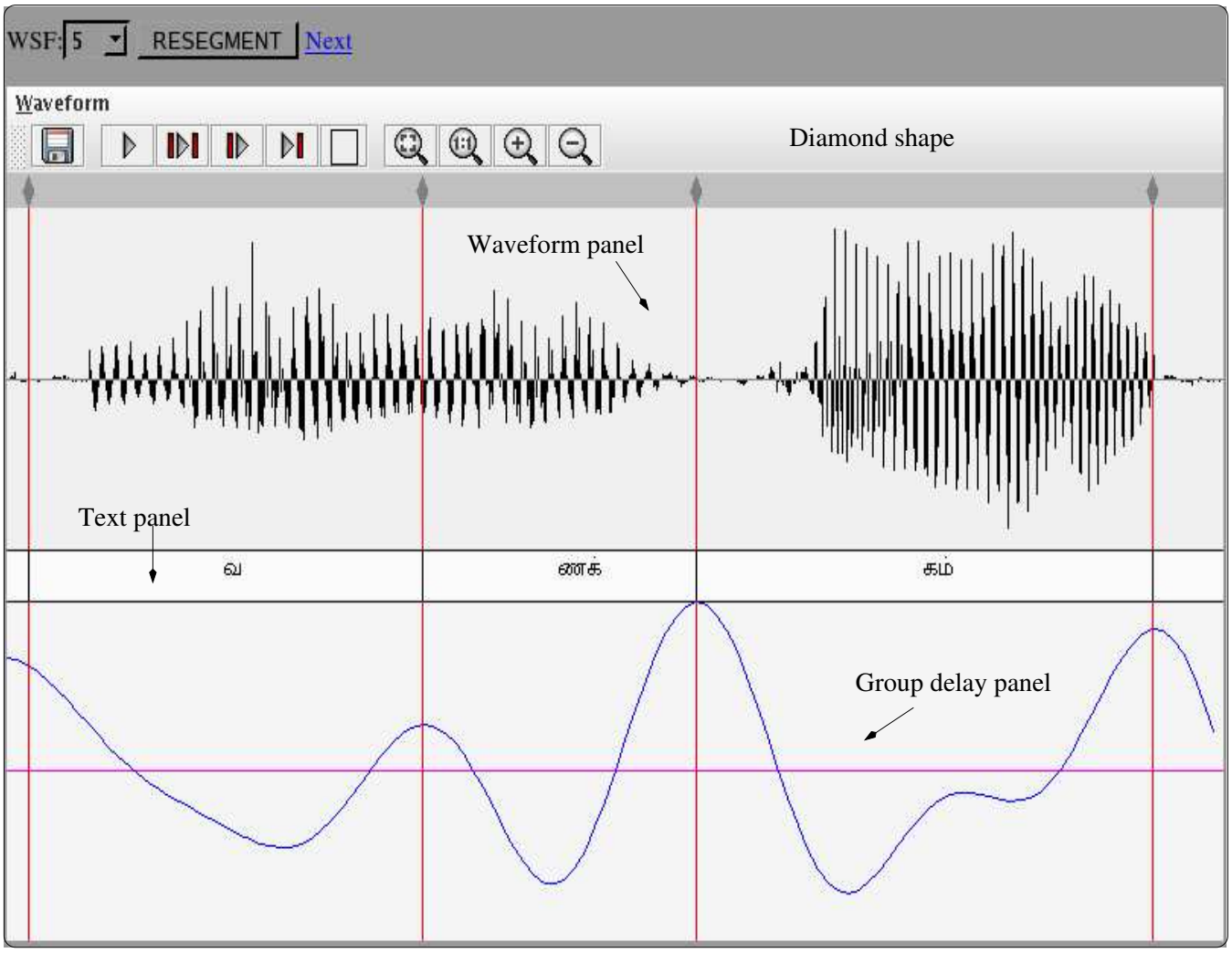

Fig. 3. A screenshot of labeling tool.

[8] T. Nagarajan, V. Kamakshi Prasad, and Hema. A. Murthy, "The minimum phase signal derived from the magnitude spectrum and its application to speech segmentation," in *Sixth Biennal Conference on Signal Processing and Communications*, pp. 95-101, Bangalore, India, July 15-18, 2001.

[9] T. Nagarajan, Hema. A. Murthy, and Rajesh. M. Hegde, "Segmentation of speech into syllable-like units," in *Proc. EUROSPEECH-03*, Geneva, Switzerland, pp. 2893-2896, Sep. 2003.

[10] Lakshmi. A, Hema. A. Murthy, "A syllable based continuous speech recognizer for Tamil," in *IEEE International Conference on Speech and Language Processing, INTERSPEECH*, pp. 1878- 1881, Pittsburgh, Pennsylvannia, September 2006.

[11] JAVA [Online]. Available: http://java.sun.com/

[12] T. Nagarajan and Hema A. Murthy, "Group delay based segmentation of spontaneous speech into syllable-like units," in *ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 115-118, Tokyo, April 2003.

[13] "Database for Indian languages," in *Speech and Vision Lab, IIT Madras*, Chennai, India, 2001.