Chapter 104
Creating Sound Glyph Database for Video Subtitling

Chitralekha Ganapati Bhat
TCS Innovation Labs, India

Sunil Kumar Kopparapu
TCS Innovation Labs, India

ABSTRACT

Accessibility of speech information in videos is a huge challenge for the hearing impaired, making a visual representation such as text subtitling essential. Unavailability of a good Automatic Speech Recognition (ASR) engine, makes automatic generation of text subtitles for resource deficient languages such as Indian languages, extremely difficult. Techniques to build such an ASR using audio and corresponding transcription in the form of broadcast news or audio books have been proposed; however, these techniques require transcriptions corresponding to the audio in editable text format, which are unavailable for resource deficient languages. In this chapter, a novel technique of building a sound-glyph database for a resource deficient language has been described. The sound-glyph database can be used effectively to subtitle videos in the same language script. Considering large volumes of data that need to be processed, we propose a parallel processing method in a multiresolution setup, harnessing the multi-core capacity of present day computers.

INTRODUCTION

Science may have found a cure for most evils; but it has found no remedy for the worst of them all - the apathy of human beings. – Helen Keller

Accessibility is one of the key design aspects for any product, to ensure that people with disabilities are able to use the product, indicates a societal growth wherein, Helen Keller's worst fears have a chance of being addressed. With increasing attention being dedicated to making any digital content accessible, text subtitling or closed captioning for videos, TV programmes, is gaining significance. Several countries have

DOI: 10.4018/978-1-5225-1759-7.ch104

mandated that all broadcasted videos be made accessible. The most common mode of making videos accessible to hearing impaired, is to provide visual cues corresponding to audio through subtitles in text format. The process of manually creating text subtitles for a video is long drawn and tedious. Alternatively, an Automatic Speech Recognition (ASR) engine can be employed to convert the audio into text and then use the text to subtitle the video, either in real-time or in the offline mode. This mechanism is efficient for resource rich languages like English. However, for resource deficient languages, especially Indian languages, this is not possible because of the absence of a good ASR in that language. This is primarily due to the non availability of a good speech corpus.

A speech corpus is a collection of speech audio files and their corresponding transcription. The sanctity of the speech corpus is measured by the quality of audio in terms of noise, accuracy of time alignment of audio and its corresponding text. Current state-of-the-art ASR technologies use audio and transcription in editable text format. There exists a wealth of open access audio and corresponding transcription in the form of news data, audio books etc. for various Indian languages. However, the transcripts of the news audio for several Indian languages are only available in non-editable form, meaning the transcripts corresponding to the audio cannot be converted into text to build a speech corpus. We propose a technique by which, using the audio and the corresponding transcripts in the image form (non-editable) to build a sound and word-glyph database. We derive a correlation between audio clips and images of the script corresponding to these audio clips by exploiting speech and image processing techniques. The central idea is to be able to build a database which represents the audio in terms of images of the script. Considering large volumes of image data that needs to be processed, we use multiresolution techniques on a multi-core processor to provide speed up in the process. The main contribution of this chapter is to build a sound-glyph database for a resource-deficient language to aid making video/audio accessible. We use multiresolution technique to reduce the size of the image and exploit inherent parallelism in the nature of the method of building the sound-glyph database.

The rest of the chapter is organized as follows, a background of the existing techniques for building a speech corpus for resource deficient language and their limitations are provided, followed by the methodology used in building the sound-glyph database using multiresolution and multi-core techniques.

Background

Through this work, we intend to address the building of a novel type of speech corpus comprising sound and its corresponding word-glyph, with special focus on Indian languages. This speech corpus is intended to be used to create video subtitles automatically. An ASR is essentially a pattern recognition engine using two types of reference models known as (a) Acoustic models and (b) Language models. These reference models or training data, are generated using a speech corpus specific to a language.

The accuracy of a state-of-the-art ASR engine is dependent on the quality and quantity of the speech corpus. A high quality speech corpus contains clean, non-noisy audio data and corresponding accurate time aligned text transcription, in addition, it has to be phonetically balanced and consist of speech spoken by diverse speakers in order to bring in an element of as much generalization as possible, for the ASR to perform optimally. For resource deficient languages, such as Indian languages, such data is not readily available. Building such a corpus from scratch is tedious and expensive.

In the absence of a good speech corpus, researchers have been looking at addressing the problem of enabling a good ASR by careful construction of language models and adaption of acoustic models. Authors outline a process of adapting language models using machine translated data from English to