Abstract

Speech recognition and understanding [1] of spontaneous speech have been a goal of research since 1970. It is a process of conversion of speech to text. The object of human speech is not just a way to convey words from one person to another but also to make the other person to understand the depth of the spoken words. Spontaneous speech is highly variable and rarely conforms to conventional assumptions and linguistically defined pronunciation rules. Specifically, there may be many different continuous speech realizations for each expertly defined phonetic unit in the dictionary. The phones may be realized in a clean and complete fashion as in read speech, or they may be realized in a sloppy and incomplete fashion as in highly spontaneous speech. For spontaneous speech, therefore, it may be beneficial to model incompletely realized variants of any phonetic unit as separate units. The purpose of this paper is to develop a method to build automatic spontaneous speech recognition for Punjabi language. Initially, the Spontaneous Speech Recognition model has been implemented for small vocabulary. We successfully build the acoustic model for Punjabi language using Sphinx4 and java programming and Language model for decoding.

Keywords: Spontaneous Speech, Punjabi Language, Synthesized Voice, ISR, CSR, Sphinx.

1. Automatic Speech Recognition System

Speech is the vocalized form of human communication. The communication with a machine requires interfaces like keyboard, mouse and screen, etc., operated with the help of software. A simple alternative to these hardware interfaces is software interface i.e. an ASR (Automatic Speech Recognition) system [2]. Automatic speech recognition is the task of taking an utterance of speech signal as an input, captured by a microphone, a telephone etc., and convert it into a text sequence as close as possible to the spoken data [4]. The main difficulties in implementation of an ASR system are due to different speaking styles of
human beings and environmental disturbances. So the main aim of an ASR system is to transform a speech signal into text message independent of the device, speaker or the surroundings in an accurate and efficient manner. Speech recognition (SR) in terms of machinery is the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words. It is a broad term which means it can recognize almost anybody’s speech, but to make the machine independent of voice, huge training data is required. There are basically two types of SR:

- Isolated speech recognition – ISR
- Continuous speech recognition – CSR

1.1 Aim of Speech Recognition System

The aim of a speech recognition system is to produce a word sequence (or possibly character sequence for languages like Mandarin) given a speech waveform. The basic structure [3] of an ASR system is shown in figure 1.1. The first stage of speech recognition is to compress the speech signals into streams of acoustic feature vectors, referred to as observations. The extracted observation vectors are assumed to contain sufficient information and be compact enough for efficient recognition. This process is known as the front-end processing or feature extraction.

**Figure 1.1 General structure of an automatic speech recognition system [3]**

Given the observation sequence, generally three main sources of information are required to recognize, or infer, the most likely word sequence: the lexicon, language model and acoustic model. The lexicon, sometimes referred to as the dictionary, is normally used in to map sub-word units, from which the acoustic models are constructed, to the actual words present in the vocabulary and language model. The language model represents the local syntactic and semantic information of the uttered sentences. It contains information about the possibility of each word sequence. The acoustic model maps the acoustic observations to the sub-word units.
1.2 Importance of Punjabi Language

Punjabi is an Indo-Aryan language spoken by 130 million (2013 estimate) [14] native speakers worldwide making it the 10th most widely spoken language in the world. It is the native language of the Punjabi people who inhabit the historical Punjab region of Pakistan and India. Amongst the Indo-Aryan languages it's unusual because it's the only tonal language. Punjabi is spoken by the majority of the population in Pakistan and 11th most widely spoken in India and the 3rd-most natively spoken language in Indian Subcontinent. Punjabi is also currently the 3rd most spoken language in the United Kingdom and the third most spoken language in Canada. Punjabi people also have significant presence in UAE, Saudi Arab, USA and Australia.

2. Spontaneous Speech Recognition

Spontaneous speech [10] and speech from written language are very different both acoustically and linguistically. Spontaneous speech includes filled pauses, repairs, hesitations, repetitions, partial words and disfluencies. Therefore, recognition of spontaneous speech will require a paradigm shift from speech to understanding where underlying messages of the speaker are extracted, instead of transcribing all the spoken words.

**Spontaneous speech:** Since user’s utterances are normally spontaneous and non-planned, and are generally characterized by disfluencies, false starts, stops in the middle and re-starts, or extra linguistic phenomena, such as cough. The speech recognition module must be able to extract, out of the speech signal, a word sequence allowing the semantic analyzer to deduce the meaning of the user’s utterance.

2.1 Model for constructing Spontaneous Speech Corpus

The appetite of today’s statistical speech [5] processing techniques for training material is well described by the aphorism: “There’s no data like more data.” Large structured collections of speech and text are essential for progress in speech recognition research. Unlike the traditional approach, in which knowledge of speech behavior is “discovered” and “documented” by human experts, statistical methods provide an automatic procedure to directly “learn” regularities in the speech data. The need for a large set of good training data is, thus, more critical than ever. However, establishing a good speech database for the computer to uncover the characteristics of the signal is not a straightforward process. There are basically two broad issues to be carefully considered: one being the content and its annotation, and the other the collecting mechanism. The recorded data needs to be verified, labeled, and annotated by people whose knowledge is introduced into the design of the
system through its learning process (i.e. via supervised training of the system after the data has been labeled) as shown in figure 1.2.

Figure 1.2 Construction of a large scale spontaneous speech corpus and processing technique

3. Literature Review

A lot of work has been done on the development of Automatic Speech Recognition systems for many languages of the world. A work ranges from the activities involved in the development and enhancement of speech corpora to the development and improvement in the speech recognition systems.

Sadaoki Furui [11] has presented the recent progress in the development of corpus-based spontaneous speech recognition technology focusing on various achievements of a Japanese 5-year national project “Spontaneous Speech: Corpus and Processing Technology”. The large-scale spontaneous speech corpus, CSJ (Corpus of Spontaneous Japanese) will be stored with XML format in a large-scale database system developed by the COE (Center of Excellence) program “Framework for Systematization and Application of Large-scale Knowledge Resources” at Tokyo Institute of Technology so that the general population can easily access and use it for research purposes. Since the recognition accuracy for spontaneous speech is still rather low, the collection of the corpus will be continued in the COE program in order to increase coverage of variations in spontaneous speech.

Monika Woszczyna [6] has presented spontaneous Thai Speech Recognition system, in this research; author expands previous work on Thai speech recognition, investigating pronunciation changes such as syllable and phoneme elisions as well as phoneme shifts in Thai spontaneous speech.
Ani Nenkova and Dan Jurafsky [2] used a richly annotated corpus of conversational speech to study the acoustic characteristics of contrastive elements and the differences between them and words at other levels of prominence. They have presented the report for automatic detection of contrastive elements based on acoustic and textual features, finding that a baseline predicting nouns and adjectives as contrastive performs on par with the best combination of features. They achieved a much better performance in a modified task of detecting contrastive elements among words that are predicted to bear pitch accent.

Hansjorg Hofmann and Sakriani Sakti [4] have attempted to recover the original word sequence from the spontaneous phoneme sequence by applying a joint sequence pronunciation model. Hereby, the whole word sequence and its effect on the alternation of the phonemes will be taken into consideration. Moreover, the system not only learns the phoneme transformation but also the mapping from the phoneme to the word directly. In this preliminary study, first the phonemes will be recognized with the present recognition system and afterwards the pronunciation variation model based on the joint-sequence approach will map from the phoneme to the word level. Our experiments use Buckeye as spontaneous speech corpus. The results show that the proposed method improves the word accuracy consistently over the conventional recognition system. The most improved system achieves up to 12.1% relative improvement to the baseline speech recognition.

Peter Mihajlik and Zoltan Tuske [8] described various morphological and acoustic modeling techniques which are evaluated on a less resourced, spontaneous Hungarian large-vocabulary continuous speech recognition (LVCSR) task. Among morphologically rich languages, Hungarian is known for its agglutinative, inflective nature that increases the data sparseness caused by a relatively small training database. Although Hungarian spelling is considered as simple phonological, a large part of the corpus is covered by words pronounced in multiple, phonemically different ways. Data-driven and language specific knowledge supported vocabulary decomposition methods are investigated in combination with phoneme- and grapheme-based acoustic modeling techniques on the given task. Word baseline and morph-based advanced baseline results are significantly outperformed by using both statistical and grammatical vocabulary decomposition methods.

Ghai and Singh [12] have developed an automatic speech recognizer to recognize continuous speech sentences by using Tri-Phone based acoustic modeling approach on HTK 3.4.1 speech engine. Performance analysis has been carried out in two phases. Overall recognition accuracy of ASR has been found to be 92.13%. In the 3rd phase, unseen data also
has been introduced in the test speech samples and its recognition has shown 90.69% accuracy. In addition, the computation time was hardly increasing.

Dua and Aggarwal [7] described the implementation of an isolated word Automatic Speech Recognition system (ASR) for an Indian regional language Punjabi. The HTK toolkit based on Hidden Markov Model (HMM), a statistical approach, is used to develop the system. Initially the system is trained for 115 distinct Punjabi words by collecting data from eight speakers and then is tested by using samples from six speakers in real time environments. To make the system more interactive and fast a GUI has been developed using JAVA platform for implementing the testing module.

Ghai and Singh [13] has developed a Phone based approach for Punjabi language speech recognition. In this work, an effort has been made to build automatic speech recognizer to recognize Isolated & Connected word speech using Phone based acoustic model approach on HTK 3.4.1 Speech Engine. An overall recognition accuracy & word error rate 92.17% and 7.83% respectively for connected word speech recognition where as overall recognition accuracy & word error rate for isolated word speech recognition are 95.00% and 5.00% respectively.

4. Speech Recognition Tools

When researchers approach the problem of core speech recognition research, they are often faced with the problem of needing to develop an entire system from scratch, even if they only want to explore one facet of the field. Open source speech recognition systems are available, such as HTK, ISIP, AVCSR [8] and earlier versions of the Sphinx systems. The available systems are typically optimized for a single approach to speech system design. As a result, these systems intrinsically create barriers to future research that departs from the original purpose of the system. First and foremost, Sphinx-4 is a modular and pluggable framework that incorporates design patterns from existing systems, with sufficient flexibility to support emerging areas of research interest. The framework is modular in that it comprises separable components dedicated to specific tasks, and it is pluggable in that modules which can be easily replaced at run time. To exercise the framework, and to provide researchers with a working system, Sphinx-4 also includes a variety of modules that implement state-of-the-art speech recognition techniques.
5. Objective of the proposed Work

So far no work has been done for spontaneous speech recognition of Punjabi language. The main area of focus of the proposed research will be on the development of an ASR system for spontaneous speech of Punjabi Language. However, recognition accuracy drastically decreases for spontaneous speech. This decrease is due to the fact that the acoustic and linguistic models used have generally been built using written language or speech from written language. Unfortunately spontaneous speech and speech from written language are very different both acoustically and linguistically. Objective of the work includes following issues:

- To develop a speech corpus for spontaneous Punjabi speech.
- To utilize the speech corpora to develop a speaker specific automatic spontaneous speech recognition system.
- To analyze the proposed system for accuracy based on standard parameters for speech recognition.

A paradigm shift from speech recognition to understanding, where the underlying messages of the speaker, i.e., meaning/content that the speaker intended to convey, are extracted, instead of simply transcribing all the spoken words, will be indispensable.

6. Implementation

Till now the Acoustic model for Punjabi language using Sphinx Toolkit has been created. At present, the system has been trained for small vocabulary. The training of the system has been done for single word of Punjabi language, two words Punjabi Language and for one sentence of Punjabi Language.

6.1 Steps for Training the Acoustic Model for Punjabi Language

To train the system for Punjabi Language, we need to require following configuration files:

6.1.1 Dictionary (Independent words are store in it): Dictionary file which maps every word to a sequence of sound units, to derive the sequence of sound units associated with each signal. Thus, in addition to the speech signals, we will also be given a set of transcripts for the database (in a single file) and two dictionaries, one in which legitimate words in the language are mapped sequences of sound units (or sub-word units), and another in which
non-speech sounds are mapped to corresponding non-speech or speech-like sound units. We will refer to the former as the language dictionary and the latter as the filler dictionary.

If the training data are:

The dictionary file (Punjabi.dic file) will look like as:

6.1.2 Filler and noise (Rejected noise are stored in it): It is also type of dictionary. Where we write the word which we want to ignore or reject while recognition.

For example:

<s>          SIL
</s>          SIL
<sil>        SIL

6.1.3 Phone: Phone file is a record of individual sound unit that we need to make a word. It is not required to write all the Punjabi characters here it actually representing the different sound unit that we have in Punjabi language. It is the most difficult part of acoustic model because; we need to have complete knowledge of Punjabi language and different sounds of Punjabi alphabets. So this will take time to fully develop. At present we identify below sounds which might optimized along with research:
6.1.4 Transcript and Fields (path of wav files): Transcript (conversation of Wav File): Transcription file (Punjabi_train.transcription and Punjabi_test.transcription) is a text file listing the transcription for each audio file. For example, in our Punjabi corpus the transcription file for test audio is like as:

<s>ਮਾਢ ਭੀਲ ਉੜਾਣ ਲੱਗ ਲਾਗ</s> (test1.Wav)

<s>ਉੜਾਣ ਲੱਗ ਲਾਗ ਮਾਢ ਭੀਲ</s> (test2.Wav)

<s>ਭੀਲ ਉੜਾਣ ਲੱਗ ਲਾਗ ਮਾਢ</s> (test3.Wav)

<s>ਉੜਾਣ ਲਾਗ ਭੀਲ ਉੜਾਣ ਲਾਗ</s> (test4.Wav)

<s>ਮਾਢ ਭੀਲ ਉੜਾਣ ਲੱਗ ਉੜਾਣ ਲਾਗ</s> (test5.Wav)

<s>ਮਾਢ ਭੀਲ ਉੜਾਣ ਲੱਗ ਉੜਾਣ ਲਾਗ</s> (test6.Wav)

<s>ਮਾਢ ਭੀਲ ਉੜਾਣ ਲੱਗ ਉੜਾਣ ਲਾਗ</s> (test7.Wav)

It's important that each line starts with <s> and ends with </s> followed by id in parentheses. Also note that parenthesis contains only the file, without speaker_n directory. It's critical to have exact match between field file and the transcription file.

We have two kinds of transcript and field files
For training purpose (Punjabi_train.transcript and Punjabi_train.fields)

For Testing Purpose (Punjabi_test.transcript and Punjabi_test.fields)

Training files are used to create feature vector which will be used later for recognition. Testing files are used by decoder to check the recognition.

6.1.5 Sphinx_train.test file: Sphinx_train.test file: This is the configuration file where all required files are present (for example field, transcript, phone, dic etc). It also defined sound formats, the frequency of sound formats and path of language model that will be used for decoding.

6.1.6 Language Model: Language model is used for decoding purpose. There are two types of models that describe language - grammars and statistical language models. Grammars describe very simple types of languages for command and control, and they are usually written by hand or generated automatically with plain code. There are many ways to build the statistical language models. When our data set is large, there is sense to use CMU language modelling toolkit. Initially, started with the grammar model, but it cannot be use for continuous and spontaneous sounds, it only be used for isolated word. But has high performance. So, the next step will be to use other statistical language model for spontaneous speech and working on to improve performance.

7. Training Process for Punjabi Language

7.1 Use the Sphinx train and Sphinx base APIs

7.2 These APIs having executable perl script. To execute this perl script, we have to use Active Python and Active Perl.

7.3 To the training purpose, enter into sphinx train folder and right a command:

```
sphinxtrain -t Punjabi setup
```

This will create etc folder and sphinx configuration files for us.

7.4 Create another folder name as Wav and place all the test sounds or trained sounds in this the wav folder.

7.5 In etc folder, we manually need to create phone, transcript, fields, dic, files. All these files are explained above.

7.6 Use these test sounds and transcript files to create feature vector.

7.7 Use the following command to create the feature vector:

```
sphinxtrain run
```
This will create a folder structure like

![Folder Structure](image)

**Figure 7.1 shows the folder structure of the Sphinx Model**

7.8 Feature vector is present in folder whose name is model_parameters

![Folder Structure](image)

**Figure 7.2 shows the internal structure of model_parameter folder**

So we use this feature vector in our application for recognition purpose.

7.9 While training it use decoder to test the training and generate log files of decoding

![Decoder Output](image)

**Figure 7.3 shows the output of decoder**
Initially, we have tested with total sentences and 42 words of Punjabi language with grammar model. Out of 42 words, only one word is failed and 2 sentences are failed out of seven sentences while decoding it.

8. Running Environment of Spontaneous Speech Model for Punjabi Language

Initially, we have created two java files for Speech Recognition of Punjabi Language.

- PunjabiLiveSpeechTest.java
- PunjabiWaveTest.java

The PunjabiLiveSpeechTest.java file will take the live test speech as input and recognize with the training audio. In PunjabiWaveTest.java we have already stored the training speech and test speech in the Wav folder.

9. Conclusion and Future work

The main objective of the research is to develop a speaker specific spontaneous speech recognition system for Punjabi speech. Therefore, first goal of this paper to develop a phonetically rich and balanced sentence based rich text corpus for Punjabi providing context based phonetic (with tri-phoneme as a phonetic context unit) cover for Punjabi speech. In future, the system will be trained for large vocabulary so that recognition rate and spontaneous speech detection can be improved for different person’s voice. The system will be tested for users with different accents in order to check and improved error rate. The Language model will be improved for proposed research work so that decoding become faster and error rates can be decreased.
References

[1] Achim Sixtus, Sirko Molau; Recent improvements of the rwth large vocabulary speech recognition system on spontaneous speech in University of Technology, 52056 Aachen, Germany, 2000, 1-6 (2000).


[10] Sadaoki Furui; Recent Advances in Spontaneous Speech Recognition and Understanding. Tokyo Institute of Technology, Department of Computer Science, 1-6 (2001).


