International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

HMM Based Distributed Speech Processing Distributed Speech Recognition

Gauri A. Deshpande[1], Pallavi S. Deshpande[2]

Abstract: ***Distributed speech recognition (DSR) is a concept which performs speech recognition over a network. It mainly focuses to assist physically challenged and visually impaired individuals to use internet services with as ease as that of others. DSR adopts the client/server methodology for implementing recognition process over internet. In combination with the HMM based recognition systems; distributed speech recognition (DSR) can deliver a highly flexible and scalable system of speech recognition. An example based on DSR can be thought of as in a meeting or discussion room, the meeting notes can be directly taken down in our handset and can be shared with all the stakeholders immediately after the meeting. The performance of speech recognition systems receiving speech that has been transmitted over mobile channels can be significantly degraded when compared to using an unmodified signal. The degradations are as a result of both the low bit rate speech coding and channel transmission errors. A Distributed Speech Recognition (DSR) system overcomes these problems by eliminating the speech channel and instead using an error protected data channel to send a parameterized representation of the speech, which is suitable for recognition. The processing is distributed between the terminal and the network. The terminal performs the feature parameter extraction, or the front-end of the speech recognition system. These features are transmitted over a data channel to a remote "back-end" recognizer. The backend server recognizes speech from the received stream of parameter vectors, uses phoneme based HMM models and recognizes the speech.***

Keywords: Distributed Speech Processing; Speech Recognition; Client-Server model; HMM; Acoustic Model

1. Introduction

The speech recognizer is implemented using the DSR methodology where the speech recognition processing is split into the client-based DSR front-end feature extraction and the server-based DSR back-end recognition. Speech features are transmitted from the DSR client to the DSR server. Subsequently, the output in the form of either the best result or an N-best list is sent from the remote DSR server back to the client. In order to enable the configuration of the DSR back-end recognizer commands are transmitted back and forth between the DSR client and the DSR server.

Another application as an example can be thought of as an auto-query response system. User queries the system and thus provides speech input. At client end, the query (speech) gets converted into feature vectors. The feature vectors are transmitted to server over a secured network. Server recognizes the feature vectors and converts them into text. This text is provided as an input to Information retrieval server. The IR server carries the database of queries and their solution. The IR server consists of a server logic module that receives the user's query in text form and determines whether the query should be sent to the search engine or can be answered directly. Once the correct solution is obtained, corresponding speech file (.wav file) will be played to answer the query.

2. Literature Survey

A. Spoken Query System

With the increasing availability of shopping sites, the count of on-line shoppers has increased drastically. Shoppers prefer to browse the items they prefer to buy. They get a choice of selecting brands, price range and a variety of categories. With the availability of speech plug-in, where the shoppers can interact, can query for the prices and select verbally will give a feel of virtual shop. It is also expected that the use of voice-based systems will increase the universe of persons willing to engage in e-commerce, e-learning etc. Application such as, various commercial programs sold by IBM (VIAVOICE) and Kurzweil (DRAGON) permit some user control of the interface (Opening and closing files) and searching but they do not present a flexible solution that can be used by a number of users across multiple cultures and without time consuming voice training.

B. Efficiency Improvement with DSR

One of the major issues which still persist in voice based systems is efficiency. Many companies are now offering technical support over the Internet [2] and some even offer live operator assistance for such queries. While this is very advantageous, it is also extremely costly and inefficient, because a person must be employed to handle such queries. This presents a practical limit those results in long wait times for responses or high labor overheads.

C. Model Used

Acoustic Model: Acoustic model defines the way speech is spoken or to be precise, the ways in which phonemes are spoken. Every phoneme is represented with an HMM (Hidden Markov) model.
Language Model: Language model defines the way words are concatenated in order to create a sentence of that model.
A typical speech recognizer consists of two distinct components [6]-

- A feature extractor (At client side)
- A pattern recognizer (At server side)- It makes use of acoustic models, language models and other domain specific knowledge.

D. Different Recognition Algorithms Used

Natural language processing is concerned with the parsing, understanding and indexing of transcribed utterances and larger linguistics units [4]. Because spontaneous speech contains many surface phenomena such as disfluencies, hesitations, repairs and restarts. Discourse markers such as

'well' and other elements which cannot be handled by the typical speech recognizer, it is the problem and the source of the large gap that separates speech recognition and natural language processing technologies. Except for silence between utterances, another problem is the absence of any marked punctuation available for segmenting the speech input into meaningful units such as utterances. For optimal NLP performance, these types of phenomena should be annotated at its input.

In speaker dependent speech recognition system, interface is trained with the user's voice which takes a lot of time and is thus very undesirable from the perspective of a WWW environment. A user may interact only a few times with a particular website. Furthermore, speaker dependent systems usually require a large user dictionary (one for each unique user) which reduces the speed of recognition. This makes it much harder to implement a real time dialog interface with satisfactory response capability. i.e. something that mirrors normal conversation on the order of 3-5 seconds is probably ideal.

In a typical language understanding system, there is typically a parser that precedes the semantic unit. Although the parser can build a hierarchical structure that spans a single sentence, parser are seldom used to build up the hierarchical structure of utterances or text that spans multiple sentences. The syntactic marking that guides parsing inside a sentence is either a weak or absent in a typical discourse [10]. So for a dialog based system that expects to have smooth conversational features, the emphasis of the semantic decoder is not only on building deeper meaning structures for the shallow analyses constructed by the parser, but also on integrating the meanings of the multiple sentences that constitute the dialog.

Until now there are two new research paths taken in deep semantic understanding of language [10]: informational and intentional. In the informational approach, the focus is on the meaning that comes from the semantic relationships between the utterance level propositions (e.g. effect cause, conditions) whereas with the intentional approach, the focus is on recognizing the intentions of the speaker (e.g. inform, request, propose).

3. Client Server Model

a) Client architecture

The advanced front-end client-side module extracts noise robust Mel-Frequency Cepstral Coefficient (MFCC) features which together with Voice Activity Detection (VAD) information are encoded sequentially and packed into speech packages for network transmission.

Voice activity detection (VAD), also known as speech activity detection or speech detection, is a technique used in speech processing in which the presence or absence of human speech is detected. The main uses of VAD are in speech coding and speech recognition. It can facilitate speech processing, and can also be used to deactivate some processes during non-speech section of an audio session: it can avoid unnecessary coding/transmission of silence packets in communication applications, saving on computation and on network bandwidth. Mel frequency cepstral coefficients are extracted at client end. These are the speech features which provide a high level of accuracy due to its closeness with the mechanism in which human perceive voice.

b) Server Architecture

DSR server receives the packets over a communication interface into a queue. Queue manages multiple packets coming to the server. The servers can be configured as SQL database servers. Queue provides the information to decoder to decode the packet information. Decoder decodes the information and passes it on to Recognition Processor. Recognition processor, performs three algorithms –

- Isolated word recognition
- Grammar based recognition
- Large Vocabulary Recognition

IR server also receives the packets over communication channel into a queue. Here also queue manages multiple packets coming to the server. Queue provides the packets to query handler. Query handler checks if it can answer the query on its own or it is required to use a search engine. Accordingly it provides the answer to the query. The solution is floating over the network again, and can be received by client and published over speaker.

4. Algorithms Used

The implementation phase starts with capturing the speech signal from Microphone and storing it into digitized format. In every phase some or the algorithm is used to change the format of data being processed. Below are the algorithms used to implement auto query answering mechanism based on DSR:

a) HMM Algorithm

As per Markov's assumption, the output of any state depends only on its previous state. With this assumption the hidden states are recognized with the consideration that the observation sequence is known. The required speech is mapped to the hidden sequence of states. This hidden sequence has to be detected. To use this model, first of all the HMM model needs to be trained for individual words or phonemes. Such phonemes are also called as senones. The phoneme HMM models can be used across the words or for multiple words. Phoneme based HMM model helps in saving the efforts of generating a huge speech corpus for covering all the possible words those a speaker can speak.

b) MFCC Algorithm

- The speech is first pre-emphasized with a pre-emphasis filter to spectrally flatten the signal.
- Then the pre-emphasized speech is separated into short segments called frame. A frame can be seen as the result of the speech waveform multiplies a rectangular pulse whose width is equal to the frame length. This will introduce significant high frequency noise at the beginning and end points of the frame because of the sudden changes from zero to signal and from signal to zero. To reduce this edge effect, an 80-points non-overlapping Hamming window is applied to each frame.

- After the FFT block, the spectrum of each frame is filtered by a set of filters, and the power of each band is calculated. To obtain a good frequency resolution, a 128-point FFT is used. Because of the symmetry property of FFT, we only need to calculate the first 64 coefficients. The filter bank consists of 33 triangular shaped band-pass filters, which ar**e** centred on equally spaced frequencies in the Mel domain between 0Hz and 4 kHz.
- We can calculate the Mel-Frequency Cepstrum from the output power of the filter bank.

c) Recognition Algorithm

1N-gram models can be imagined as placing a small window over a sentence or a text, in which only n words are visible at the same time. The simplest n-gram model is therefore a so-called unigram model. This is a model in which we only look at one word at a time. The sentence Colorless green ideas sleep furiously, for instance, contains five unigrams: "colorless", "green", "ideas", "sleep", and "furiously". Of course, this is not very informative, as these are just the words that form the sentence. In fact, N-grams start to become interesting when n is two (a bigram) or greater.

To stick to our 'window' analogy, we could say that all bigrams of a sentence can be found by placing a window on its first two words, and by moving this window to the right one word at a time in a stepwise manner. We then repeat this procedure, until the window covers the last two words of a sentence. In fact, the same holds for unigrams and N-grams with n greater than two. So, say we have a body of text represented as a list of words or tokens.

d) Information Retrieval Algorithm

The domain and knowledge manager is the main component of this section. It receives the query in a structured representation (for instance, KIF or CGIF), [8] and then performs one of the following tasks based on the domain knowledge and the refinement information. It verifies the correctness of the query. For instance, it identifies whether the query is pertinent to the domain and possibly decides if it is feasible to answer.

It enriches the query. Basically, it considers the following actions:

1) Transforming modern terms into out-dated terms.
2) Separating the initial request into several more precise queries
3) Adding information such as dates or names to the query in order to restrict the searching space.
4) Selects the most adequate searching engine, depending on the kind of query. One can request historical information or ask for an explanation about anachronism terms. Accordingly, we are considering engines for:
   - Word resolution
   - Document searching
   - Question answering

5. Conclusion

The system adopts a distributed architecture in which the speech recognizer and the knowledge-based IR system are located in different servers. The spoken queries are processed using the DSR technology. With this technology, there is a great help provided to the society by enabling the vision impaired people from successfully using the internet services to the fullest. Future work will consider using *n*-gram language models in the DSR server. It also needs implementation of HMM and MFCC algorithms. Along with IR methodology to provide real time correct responses for the queries raised by user. Post this implementation; the system can be designed for N number of clients interacting with the DSR and IR servers. There can be a prototype defined for the network over which all clients communicate with DSR server and IR server.

References

[1] WeiQi Zhang, Archit. Dev. Lab., Intel Corp., China ; Liang He, Yen-Lu Chow, RongZhen Yang, "The study on distributed speech recognition system", Vol 3, June 2000
[2] Junhui Zhao, Xiang Xie, Jingming Kuang, "The Performance Evaluation of Distributed Speech Recognition for Chinese Digits", *Proceedings of the17 th International Conference on Advanced Information Networking and Applications (AINA'03), IEEE Computer Society*, March 2003
[3] Wenjing Han ; Inst. for Human-Machine Commun., Tech. Univ. Munchen, München, Germany ; Zixing Zhang ; Jun Deng ; Wollmer, M, "Towards distributed recognition of emotion from speech", *IEEE Conference Publication*, Vol 3, June 2013
[4] Ian Bennett, Palo Alto, CA (US), "Method For Processing Speech Data For a Distributed Recognition System" *United State Patent, US 7672841*, March 2010.
[5] Alexander Sorin, Haifa (IL), "Restoration of High-Order Mel Frequency Cepstral Coefficients", *United States Patent, US 20090144058*, June 2009
[6] Naveen Srinivasamurthy, Antonio Ortega, Shrikanth Narayanan, "Efficient scalable encoding for distributed speech recognition", *United States Department of Electrical Engineering-Systems, Signal and Image Processing Institute, Integrated Media Systems Center*, November 2005.
[7] Tom Brøndsted1, Henrik Legind Larsen2, Lars Bo Larsen1, Børge Lindberg1, Daniel Ortiz-Arroyo2, Zheng-Hua Tan1, Haitian Xu1, "Mobile Information Access with Spoken Query Answering", *Department of Communication Technology 2 Software Intelligence and* Security Research Center (SIS-RC), Esbjerg Aalborg University, Denmark, December 2012.
[8] J A Gonzalez, A Lopez-Lopez, J Munoz-Arteaga, M Montez, Y Gomez, "Natural Language Dialogue System for Information Retrieval", Institute National Astrophysics, May 2010.
[9] M. H. Moattar and M. M. Homayounpour, "A simple but efficient real-time voice activity detection algorithm", 17th European signal processing conference, August 2009.
[10] Ian Benett, "Natural language speech lattice containing semantic variants", *United States Patent, US 7873519 B2*, January 2011