An Overview of Distributed Speech Recognition over WMN

Jyoti Prakash Vengurlekar
Ramrao Adik Institute of Technology
Nerul, Navi Mumbai
vengurlekar.jyoti13@gmail.com

Krupali Suresh Raut
Ramrao Adik Institute of Technology
Nerul, Navi Mumbai
raut.krupali7@gmail.com

Shital Mali
Ramrao Adik Institute of Technology
Nerul, Navi Mumbai
shital_khalane@rediffmail.com

Abstract— In this paper we have discussed method of speech recognition over wireless mesh networks. The typical distributed speech recognition system the processing is distributed between the client & server. System that follows ETSI standards performs feature extraction at the client. The features are then compressed and transmitted to the server over dedicated channel where they are decoded and delivered to the speech processing back end which generally uses a statistical modeling method. his new framework of packet switched network (WMN) has an additional advantage of supporting novel applications which need to handle large volumes of speech data over WMNs. For distributed speech processing conventional ETSI is used then WMN is introduced and then router aided distributed speech recognition is proposed.

Index Terms— Distributed speech recognition, WMN, MFCC, MVDR, GMM.

1 INTRODUCTION

Distributed speech recognition (DSR) technology dramatically improves recognition performance, while minimizing the memory and CPU requirements on the device. This is achieved by using a noise robust front-end and by eliminating the detrimental effects of low bit-rate. DSR standard front-end increases accuracy in speech recognition. It is based on a data network & fits into the wireless Internet architecture due to the new standards in wireless application protocol (WAP). It is attractive since it focuses on speech recognition & multi-modal applications. Standards in this area are produced to work well with modern speech recognition systems. Also standards are implemented to minimize the impact of bit errors on standard communication channels. It is integrated in the data network which makes easy to envision integrating authentication with Internet security.

To enable widespread applications using DSR in the market place, a standard for the front-end is needed to ensure compatibility between the terminal and the remote recognizer. The Aurora DSR Working Group within ETSI has been actively developing this standard over the last two years. The first DSR standard was published by ETSI in February 2000.

2 WMN

Wireless mesh networks (WMNs) have emerged as a key technology for next-generation wireless networking. Because of their advantages over other wireless networks, WMNs are undergoing rapid progress and inspiring numerous applications. However, many technical issues still exist in this field.

Wireless mesh networks (WMNs) are dynamically self-organized and self-configured, with the nodes in the network automatically establishing an ad hoc network and maintaining the mesh connectivity. WMNs are comprised of two types of nodes: mesh routers and mesh clients. Mesh routers are advantageous over convential router in terms routing functions to support mesh networking, large coverage with lower transmission power. WMNs diversify the capabilities of ad-hoc networks in terms of low up-front cost, easy network maintenance, robustness, reliable service coverage, etc. Therefore, in addition to being widely accepted in the traditional application sectors of ad hoc networks, WMNs are undergoing rapid commercialization in many other application scenarios such as broadband home networking, community networking, building automation, high speed metropolitan area networks, and enterprise networking [2].

The network architecture of WMNs can be classified into three types wiz. infrastructure/ backbone architecture, Client WMNs, hybrid WMN [2]. Hybrid WMN is preferred over other architecture as it supports both mesh clients & mesh routers. The figure is shown below for illustration.

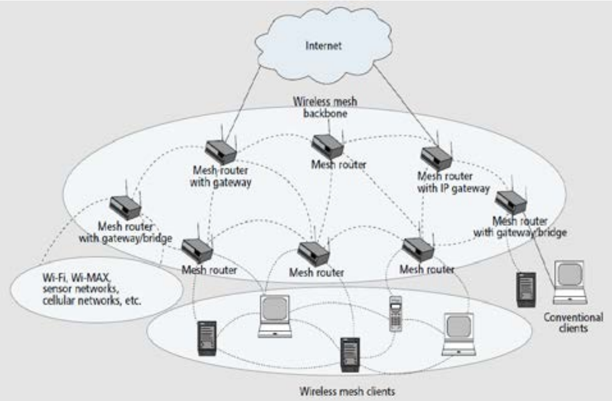
Internet
Wireless mesh backbone
Mesh router with gateway
Mesh router
Mesh router with IP gateway
Mesh router with gateway/bridge
Mesh router with gateway/bridge
Mesh router
Mesh router
Mesh router
Wi-Fi, Wi-MAX, sensor networks, cellular networks, etc.
Conventional clients
Wireless mesh clients

Fig. 1: Hybrid WMN.

This architecture is the combination of infrastructure and client meshing as shown in Fig.1. Mesh clients can access the network through mesh routers as well as directly meshing with other mesh clients. While the infrastructure provides connectivity to other networks such as the Internet, Wi-Fi, cellular, and sensor networks, the routing capabilities of clients provide improved connectivity and coverage inside WMNs.

3 A FRAMEWORK FOR DSR WMN

3.1 Basic DSR system

DSR works by splitting the processing required for speech recognition between the device and network servers, instead of sending the speech data to the server and having all the processing done there. Beginning the processing on the device, or 'front-end', enables the device itself to extract spectral features from the speech. These features are compressed, error protected, and transmitted over the wireless channel to the server, or 'back-end'. Once the compressed features have arrived at the server, the server can then convert the incoming stream of features into text.

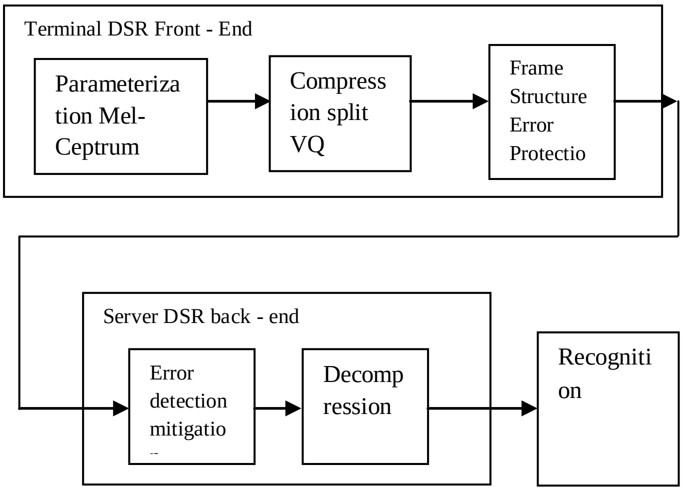

Fig. 2: Basic DSR system.

3.2 Framework of DSR over WMN

The ETSI standard moves the front end of the speech recognition process to the client while the more resource intensive back end processing is done at the server. The standard also lays down rules for speech feature compression and error control coding at the client. It also defines the rules for decoding and error mitigation at the server prior to the back end processing. To improve the speech recognition performance multiple feature streams can be used. This form of distributed multi stream processing acquires speech data from various devices connected to the network having heterogeneous processing capacities, extracts multiple features and performs recognition at the different WMN nodes which form the WMN.

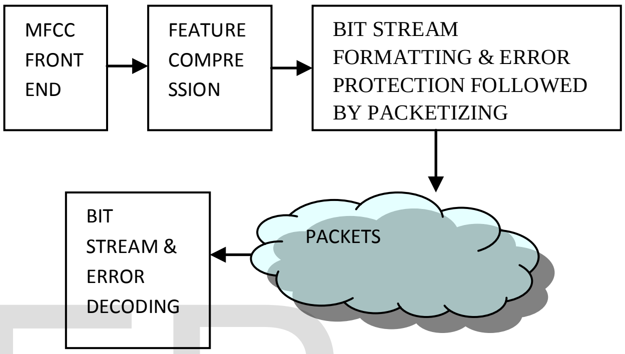
MFCC FRONT END
FEATURE COMPRESSION
BIT STREAM FORMATTING & ERROR PROTECTION FOLLOWED BY PACKETIZING
PACKETS
BIT STREAM & ERROR DECODING

Fig. 3: A Framework for DSR over WMN.

Speech recognition is a special case of pattern recognition. There are two phases in pattern recognition viz. are training & testing. The process of extraction is common for both the phases. During the training phase the parameters of classification model is estimated using large number of class examples (training data). During the testing or recognition phase tested data is matched with the training model of each and every class.

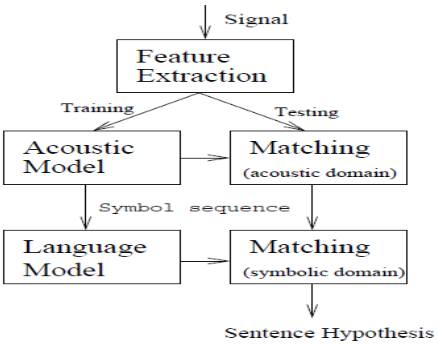

Fig. 4: Typical speech recognition system.

The recognition takes place in the two domains, acoustic and symbolic. In the acoustic domain feature vector related to small segment of test speech is matched with the acoustic model of each and every class. The segment is assigned with the label of the class with highest matching score. This process is repeated for every feature vector in feature vector sequence from the test data. The resultant sequence of label is processed in conjunction with the language model to give recognized sentence [15]. Acoustic model in the speech recognition should be capable of modeling predictable variations (context dependent variations) of acoustic characteristic of speech sound as well as other variations due to speaker. HMM is the best suited model [12]. The language model is used to derive best sentence hypothesis subject to the constraint of the language. It incorporates various types of linguistic information [12].

3.3 DSR FRONT END STANDARDS

Figure 5, shows a detailed block diagram of the processing stages for the DSR front end. At the terminal the speech signal is sampled and parameterized using a mel-cepstrum algorithm to generate 12 cepstral coefficients together with C0 and a log energy parameter. These are then compressed to obtain a lower data rate for transmission. To be suitable for today's wireless networks a data rate of 4800 b/s was chosen as the requirement. The compressed parameters are formatted into a defined bit stream for transmission. It is anticipated that the DSR bit stream will be used as a payload in other higher level protocols when deployed in specific systems supporting DSR applications. Thus the standard does not cover the areas of data transmission or any higher level application protocols that may run over them. In this respect it is similar to speech codec standards where the codec is specified separately to the systems that use it [11]. The mel-cepstrum was chosen as the feature set for the first standard because of its widespread use throughout the speech recognition industry [11].

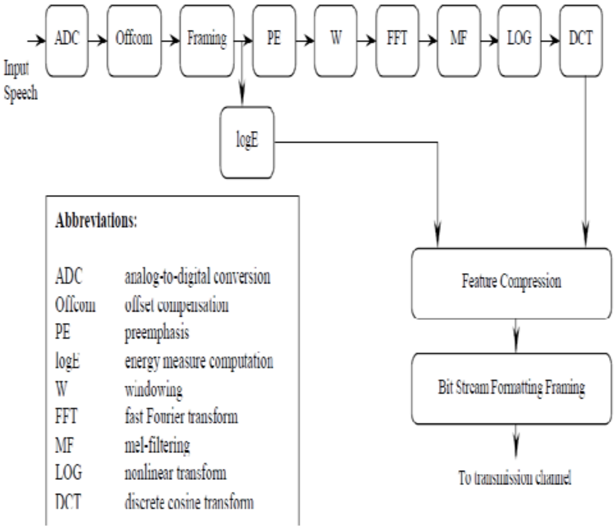
Input
Speech
ADC
Offcom
Framing
PE
W
FFT
MF
LOG
DCT
logE
Abbreviations:
ADC analog-to-digital conversion
Offcom offset compensation
PE preemphasis
logE energy measure computation
W windowing
FFT fast Fourier transform
MF mel-filtering
LOG nonlinear transform
DCT discrete cosine transform
Feature Compression
Bit Stream Formatting Framing
To transmission channel

Fig. 5: DSR front end.

3.4 DSR over Wireless Channel

There are several alternative architectures for applications incorporating speech recognition technology on the WWW three of which are listed here [3].

A. Server-Only Processing

B. Client-Only Processing

C. C. Client-Server Processing

The communication channels between the client and the server may have limited bandwidth. That would be a realistic assumption in applications that communicate over the Internet or through wireless channels. The architecture [3] of the client-server speech recognition is shown in Figure 6.

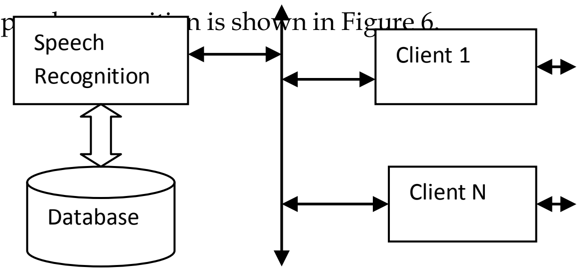

Fig. 6. Client-server speech recognition system.

A central server provides speech recognition services. The clients are deployed on heterogeneous environments, such as personal computers, smart devices, and mobile devices. Speech is captured by the clients and, after some local processing, the information is sent to the server. The server recognizes the speech according to an application framework and sends the result string or action back to the client [3].

3.5 Coding of ceptral features

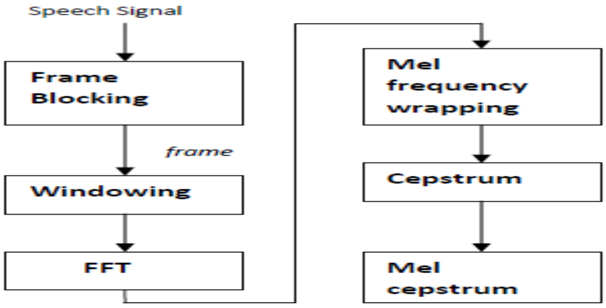

Fig. 7: MFCC block diagram.

Mel-Filtered Cepstral Coefficients (MFCC) is feature extraction set. Cepstral coefficients derived from a modified short-time spectrums the most popular feature set and has been empirically observed to be the most effective for speech recognition.

3.6 MVDR based front end

Next step in the DSR is the MVDR [4] which is robust feature extraction method for continuous speech recognition. Minimum Variance Distortion less Response (MVDR) is the method of spectrum estimation and a feature trajectory smoothing technique for reducing the variance in the feature vectors. When the above mentioned method evaluated on continuous speech recognition tasks in noisy environment gave an average relative improvement in WER of greater than 30%.

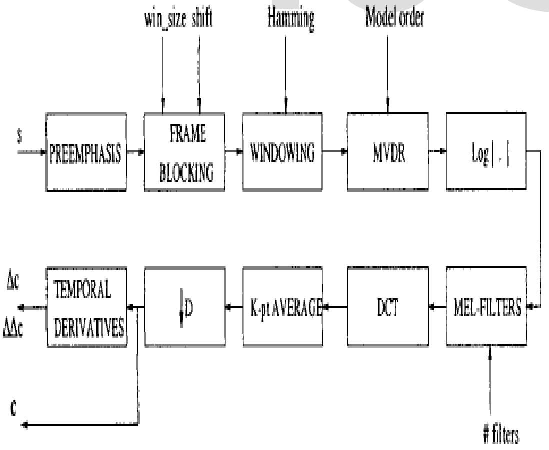

Fig. 8: Schematic diagram of the MVDR-based front-end processor.

The above figure shows a schematic diagram of the MVDR based front-end processor. Instead of generating a single MFCC vector from a frame of speech samples from the start of the current frame to the start of the next frame are split into several overlapping segments and an MFCC vector are computed from each segment. These vectors are then averaged to get the smoothed MFCC vector for that frame. This is equivalent to generating feature vectors at a high frame rate and down sampling the resulting trajectories after low pass filtering in the time domain. The filtering operation is performed by simple averaging [4].

3.7 GMM

GMM is Gaussian Mixture model used for the speaker identification. We have to recognize and classify the speeches of different persons. Estimation and Maximization algorithm is used, for finding the maximum likelihood solution for a model with latent variables, to test the later speeches against the database of all speakers who enrolled in the database. The performance of Speech identification is evaluated by speech databases TIMIT having BW of 8Khz and NTIMIT has band limiting (3300Hz) due to additional non linear distortion.

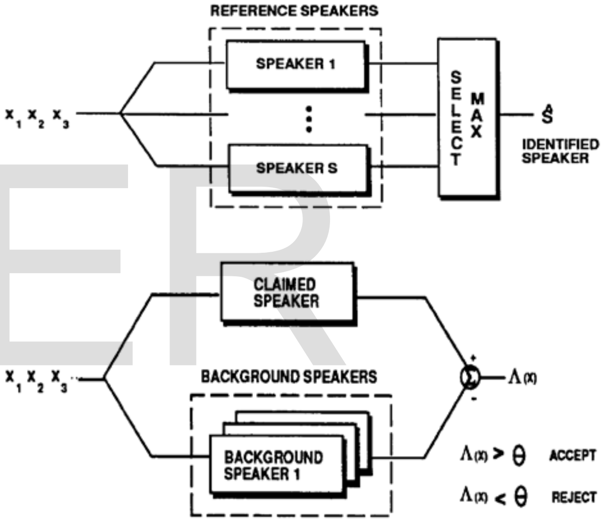

Fig. 9: Speaker recognition systems (a) Identification system (b) Verification system.

Over the last decade, the GMM is standard classifier for text-independent speaker recognition. It operates on atomic levels of speech and can be effective with very small amounts of speaker specific training data. The primary focus of this work was on a task domain for a real application, such as voice mail labeling and retrieval. The Gaussian Mixture speaker model was specifically evaluated for identification tasks using short duration utterances from unconstrained conversational speech, possibly transmitted over noisy telephone channels. The experimental evaluation examined several aspects of using Gaussian mixture speaker models for text independent speaker identification [18].

Applications of DSR include voice-activated web portals, menu browsing and voice-operated personal digital assistants. In order to provide high recognition accuracy over a wide range of channel conditions with low bit rate, delay and complexity for the client in wireless communications media.

4 CONCLUSION

In this paper, we have described a novel paradigm of distributed speech processing over WMN using multiple feature streams. The framework also provides the additional flexibility of voice users accessing any node in the network and falling back on conventional client-server distributed speech recognition. An analytical estimate shows that on an average, the bandwidth saving compared to the speech transfer is about 36%. The WMN router that faces the highest processor demand, for a particular flow may spend close to 10% of the speech processing workload required for a session. Although the idea of distributed speech processing over WMN is currently analyzing and exploring methods to implement such a system in a real deployment scenario.

REFERENCES

1. Rajesh M. Hegde and B.S. Manoj IIT Kanpur, 'Distributed speech recognition over wireless mesh networks" IEEE, 2011
2. I. F. Akyildiz, X. Wang and W. Wang, "Wireless mesh networks: A survey" Computer Networks, vol. 47(4), pp. 445-487, Mar 2005
3. V. V. Digalakis, L. G. Neumeyer and M. Perakakis," Quantization of cepstral parameters for speech recognition over the world wide web," IEEE J. select Areas Commun., vol. 17(1), pp. 82-90, Jan 1999.
4. S. Dharanipragada and B. D. Rao, "MVDR based feature extraction for robust speech recognition," in proceedings of IEEE Int. Conf. Acoust., speech and signal processing, Utah, May 2001, vol. 1, pp. 309-312
5. D. B. Johnson and D. A. Maltz, "Dynamic Source Routing in Ad Hoc Wireless Netwok", Mobile Computing, PP. 152-181, 1996
6. A. S. Tanenbum, computer networks, Prentice Hall PTR, NJ, USA, 2002
7. NTIS, The DARPA TIMIT Acoustic – Phonetic continous speech corpus, 1993
8. A. Bernarda and A. Alwan, "Low bit rate distributed Speech recognition for packet based and wireless communication, "IEEE transactions on speech and audio processing, vol. 10(8), November 2002
9. M. Soltane and N. Doghmane, N. Guersi, "State of the art: speech biometric verification", Journal of Information Technology Review, Vol. 1(3), pp. 125-133, August 2010
10. Y. S. Zhang et al., Wireless Mesh Network: Architecture, Protocols, and standards, CRC Press, 2006
11. Charles C. Broun, William M. Campbell, David Pearce, Holly Kelleher "Distributed Speaker Recognition using the ETSI Distributed speech recognition standard", Motorola Human Interface lab
12. Plannerer, "An Introduction to Speech Recognition", March 28, 2005
13. Angel M. Gomez, Antonio M. Peinado and Antonio I. Rubio, "Recognition of coded speech transmitted over wireless channels", IEEE transactions on Wireless Communication, vol. 5, No. 9, September 2006
14. Djohara Benyamina, Abdelhakim Hafid and Michel Gendreau, "Wireless Mesh Networks Design - a Survey" IEEE communication surveys & tutorials, vol. 14 No. 2, second quarter 2012
15. Samudravijaya K, "Speech and speaker recognition: A tutorial", Tata Institute of Fundamental Research
16. Tomi Kinnunen, Filip Sedlak, Johan Sandberg, Maria Hansson-Sandsten, Haizhou Li, "Low Variance Multitaper MFCC features: A case study in robust speaker verification", IEEE transactions on audio, speech and language processing, vol. 20, No. 7, September 2012
17. G. Suvarna Kumar et. al., "Speaker Recognition using GMM", International Journal of Engineering Science and Technology, Vol. 2(6), pp. 2428-2436, 2010
18. Douglas A. Reynolds, "Speaker Identification and verification using Gaussian Mixture speaker models", Elsevier, speech communication 17 (1995) 91-108
19. Dale Isaacs, Professor Daniel J. Mashao," A Tutorial on Distributed Speech Recognition for Wireless Mobile Devices", Speech Technology and Research Group (STAR)