2016 Conference of The Oriental Chapter of International Committee
for Coordination and Standardization of Speech Databases and Assessment Technique (O-COCOSDA)
26-28 October 2016, Bali, Indonesia

Urdu Speech Recognition System for District Names of Pakistan: Development, Challenges and Solutions

Muhammad Qasim, Sohaib Nawaz, Sarmad Hussain
Centre for Language Engineering, Al-Khwarizmi
Institute of Compute Sciences
UET, Lahore, Pakistan
firstname.lastname@kics.edu.pk

Tania Habib
Computer Science and Engineering Department
UET, Lahore, Pakistan
tania.habib@uet.edu.pk

***Abstract*— Speech interfaces provide people an easy and comfortable means to interact with computer systems. Speech recognition is a core component of speech interfaces which recognizes human speech in a particular language. Some small vocabulary speech recognition systems for Urdu language with reasonable accuracy have been developed but these systems are either speaker dependent or unable to handle the large accent variations that exists among people of Pakistan. This paper presents a speaker independent Urdu speech recognition system for district names of Pakistan that performs well for major accents of Urdu. Two methods have been studied and analyzed to handle accent variation. The systems have been tested in laboratory and field and their results have been reported in this paper. We conclude that accent independent system performs better for isolated words and addition of field data in the training of system improves the overall recognition accuracy.**

Keywords—speech recognition; isolated word; speaker independent; accent variation; field testing

I. Introduction

The access to online information has become essential for development in today’s age and literate population of world is getting benefit from it. But on the other hand, barriers like low literacy rate and internet connectivity are hindering illiterate and semi-literate population to access invaluable online resources. To overcome this challenge, speech interfaces are used to provide information to the users in their local languages. Speech interfaces are also helpful for visually challenged persons. The speech recognition system is a fundamental component of these speech interfaces.

Speech recognition systems are categorized into three types; isolated word, connected word and continuous speech recognition systems. Isolated word systems recognize only isolated word, connected word systems recognize connected words while continuous speech recognition systems recognize words spoken as in natural conversation. Speech recognition systems are used in spoken dialog, dictation, language learning and speech translation systems to name a few.

Speech recognition systems require the design and development of speech corpus, language models and grammar specifications related to the language for which system is to be developed. Corpus development includes the collection, careful annotation, cleaning and verification of speech data. These resources are mostly unavailable for Urdu language due to which speech recognition for Urdu language is still at a very basic level. Few speech corpora for Urdu language consisting isolated words have recently been developed that may be used to develop isolated words speech recognition systems.

This paper presents an isolated word Urdu speech recognition system to recognize 139 district names of Pakistan. The system is developed for use in a spoken dialog system that provides weather information to the citizens of Pakistan in Urdu language. In order to develop a speaker independent system, we need to handle the accent variation of speakers calling from all over Pakistan. Major languages being spoken in Pakistan include Punjabi, Sindhi, Pashto, Balochi, Seraiki and Urdu. Accent of each language’s speaker is significantly different than others. Therefore, we have implemented and evaluated different techniques to handle this accent variation. The system also needs to be robust enough to work in low to mild noise environments. This is achieved through preprocessing of speech input signal. Rigorous field testing of speech recognition system in real world environment is conducted to evaluate the performance of the system.

The rest of the paper is structured as follows: Section 2 describes literature review. Section 3 describes the methodology used to build ASR for weather forecast system and their laboratory results. Section 4 discusses the results of the system’s field testing and improvements made in the system based on the results, Finally, Section 5 concludes our discussion and results.

II. Literature Review

Speech recognition systems have been in use for decades now. The earliest speech recognition system, Audrey [1], was developed in 1952 in Bell Laboratories which recognized digits from a single-speaker. Speech recognition has advanced considerably in last few decades. Speech recognition systems for English include Nuance’s Dragon, BBN’s BYBLOS [2] and MIT’s SUMMIT [3]. The BBN continuous speech recognition system, BYBLOS [2], is a large vocabulary speech recognition system. BYBLOS is tested for two domains; Electronic Mail (EMAIL) consisting of 334 words and Naval Database Retrieval consisting of 354 words. The system achieves an accuracy of 98.5% for speaker-dependent mode and 97% for speaker-adapted mode. The SUMMIT [3] speech

978-1-5090-3516-8/16/$31.00 ©2016 IEEE

recognition system developed by MIT used 1500 sentences recorded from 300 speakers for training. This system achieved an accuracy of 87% on DARPA Resource Management Task. The study [4] discusses continuous speech recognition system for French language consisting of 65,000 words and reports an accuracy of 88.8%.

Speech recognition for resource rich languages has excelled greatly but it is still at a rudimentary level for under resource languages such as Urdu. In order to develop speech recognition systems for under resource languages, Speech-based Automated Learning of Accent and Articulation Mapping (SALAAM) [5] uses the existing available resources of “developed languages” like English to develop systems for “developing languages”.

The SALAAM method was tested for 10 diverse languages and yielded word recognition rate around 90% for 3 to 10 words. An improved method of speech recognition for low resource languages based on SALAAM is presented in [6]. They used US English as source language and target languages included Yoruba, Hindi and Hebrew. They reported word recognition rate up to 90% for vocabulary size of 20-30 words but it drops sharply for higher vocabulary size (word recognition rate of 50 word vocabulary for Hindi is around 77%). Due to relatively large vocabulary size consisting of 139 district names of Pakistan, the SALAAM method was not used in our study.

Recently, speech recognition systems for several under-resource languages have been developed. These include isolated words system for Arabic language [7]. The system is used to assists callers using speech interface with a recognition rate of 88.8%. The study [8] presents Hindi language isolated words speech recognition system consisting of 113 words. The training data was collected from five male and four female speakers. The word recognition rate for the speakers present in training data is 96.61%. For speakers not present in training data, word recognition rate is 95.49%. An isolated word speech recognizer for Bangla language has been developed that consists of 100 vocabulary words [9]. They reported an accuracy of 90% for speaker dependent system and 70% for speaker independent system.

The development of an isolated speech recognition system for Urdu language is described in [10]. The vocabulary size was 52 words and training data was collected from ten speakers. The word recognition rate for the speakers present in training data is 94.67%. For speakers not present in training data, word recognition rate is 89.34%. The system is not feasible for a real word application as the data used is statistically insignificant.

A first attempt to develop a continuous speech recognition system for Urdu language is discussed in [11]. The corpus for the system consisted of data recorded from 42 male and 40 female speakers with total duration of 45 hours. The reported word recognition rate was around 40% due to insufficient noise modeling, diacritic issues and lack of accent variation handling. In this paper, we present an isolated word speech recognizer supporting six major accents of Urdu with reasonable accuracy both in laboratory testing and field testing.

III. Methodology

This section presents the development of speaker independent Automatic Speech Recognition (ASR) system for district names of Pakistan. In order to handle accent variation, we developed two types of speech recognition system; an accent dependent system and an accent independent system. We have used CMU Sphinx [12] toolkit for the development of speech recognition systems. Additionally, we developed an accent classifier system to identify the accent of any unknown utterance. The following sub sections discuss the development of speech corpus and speech recognition systems.

A. Speech corpus

The speech corpus consisted of more than nine hours of speech data recorded from 300 speakers (both male and female) from all over Pakistan covering the aforementioned six major accents. The detail of the corpus is given in Table I. The data was collected over mobile channel using Asterisk Private Branch Exchange (PBX) at 8 KHz sampling rate and digitization rate of 16 bits. The data was cleaned and verified by expert linguists using a set of developed guidelines [13].

TABLE I. Corpus Details

First language of speaker	Number of Utterances	Duration (in minutes)
Urdu	3424	51
Punjabi	4493	71
Pashto	7934	125
Sindhi	5231	70
Balochi	13262	113
Seraiki	11494	141
Total	41293	574

B. Pre-processing

Voice Activity Detector (VAD) is pre-processing technique to separate out the voice and non-voice portions of a user input. We used procedure outlined in [14] for VAD and only the speech portion of file is fed to the ASR for recognition. This pre-processing improves the accuracy of speech recognition systems. VAD works by computing the statistics of background noise from the first 200 samples of the recorded file, therefore it can handle static noise like noise of fans in the background but fails to handle the dynamic noise like people talking in the background. VAD and background noise are closely linked and it is not incorrect to treat the errors of voice activity detection as consequences of dynamic noise. Most of the errors in voice activity detection are also due to background noise – if noise is higher and comparable in magnitude to the spoken word, it will inevitably lead to cutting some portion of the speech as well and the final speech given as input to ASR will be an incomplete word. On the other hand, if noise is very low at the start but changes midway in the recording, like someone starts speaking loudly in the background or a car horn honks, it will lead to noise being treated as speech and the ASR will most likely give wrong output.

C. *Accent Identification*

Accent is linked to the articulation pattern followed by a speaker when producing a particular sound. It is associated with the first language of speaker which affects the production of speech. In this work, we focused on identification of six main accents of Urdu i.e., Urdu, Punjabi, Sindhi, Balochi, Seraiki, and Pashto. In our previous work [15], we conducted a study to classify five of these accents based on formant frequencies and Mel Frequency Cepstral Coefficients (MFCCs). Three approaches were used to classify different accents: 1) using Support Vector Machine (SVM) and Random Forest, 2) using Gaussian Mixture Models (GMMs) and 3) using confidence scores of phone ASRs.

1) Accent classification using SVM and Random Forest: Using machine learning classification algorithms, an experiment was carried out to classify six different accents of Urdu language. For classification of accents, feature vector used was Mel Frequency Cepstral Coefficients (MFCCs). The MFCCs were computed from complete words of particular accents and then averaged over a window of few milliseconds. The values for frame length, frame shift and averaging window used in experiment were 10ms, 7ms and 50ms respectively. The kernel used was polynomial with cost value of 10, gamma and co-efficient value of 0.5. The number of trees and depth of tree for Random Forest were 400 and 0 respectively. The number of utterances used for each accent were 424. This method resulted in an accuracy of 44%.

2) Accent classification using Gaussian Mixture Models: The system was built using the Gaussian Mixture Models where each accent was represented by a mixture of Gaussian densities. In training phase, MFCC features of all the wave files of a single accent were computed and stacked together. Expectation Maximization Algorithm was used to find the optimal parameters of Gaussian Mixtures and the weight of each mixture. In testing phase, after computing MFCC feature vector for a single file, posterior probabilities of each feature vector were computed for all three accents. Each test utterance was assigned to the accent for which its posterior probability was maximum. Same number of files (3861) was used in training for each accent, but number of files in testing was different according to the data available for each accent. The overall accuracy of the system was about 61% which is the weighted average of the accuracies of the individual accents.

3) Accent classification using confidence scores of phone ASRs: This method estimates the accent of a test utterance by comparing its confidence scores after being decoded by phone ASRs of all accents. Results of using this method are better than other SVM and GMM based methods. The overall accuracy of this system is 65.71%, with Punjabi having the least accuracy and Urdu having the most. The reasons for poor accent identification may be because of the fact that vocabulary includes proper nouns and isolated words which may not adequately capture accent variations.

D. *Accent Dependent ASRs*

A separate system for each accent using the data available for each accent was developed. Sphinx toolkit was used to develop the systems. Table II shows the results of speech recognition system for each accent.

TABLE II. ACCENT DEPENDENT ASR SYSTEMS DATA AND ACCURACY

Accent	Training Utterances	Testing Utterances	Accuracy (%age)
Punjabi	3476	793	91.29
Pashto	6202	1566	93.99
Urdu	2661	616	92.37
Balochi	10757	2505	92.71
Saraiki	8998	2243	95.05
Sindhi	4301	1075	91.81
Overall AD ASRs Accuracy			92.87

The developed accent dependent system are used with accent identifier for recognition. Figure 1 shows the combination of accent identification with accent dependent ASRs.

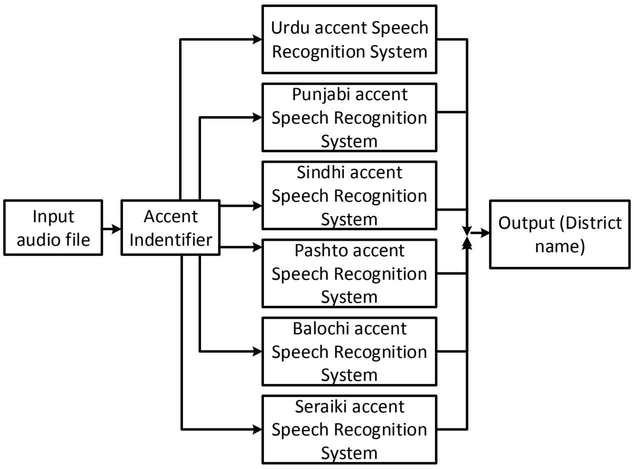
Input audio file
Accent Indentifier
Urdu accent Speech Recognition System
Punjabi accent Speech Recognition System
Sindhi accent Speech Recognition System
Pashto accent Speech Recognition System
Balochi accent Speech Recognition System
Seraiki accent Speech Recognition System
Output (District name)

Figure 1: Accent dependent ASRs with accent classifier.

E. *Accent Independent ASR*

The accent independent speech recognition system for weather domain is built using the speech data of Punjabi, Urdu, Sindhi, Pashto, Balochi and Seraiki speakers. The training data consisted of 80% of the corpus for each accent while rest of the 20% of each accent was used for testing the system. The number of training utterances was 28805 while 7759 utterances were used for testing. The recognition accuracy of the system was 91.98%.

IV. Field Testing

The aim of the field testing is to compute accuracy of ASR system in real-world conditions. Based on the amount of noise present in the surroundings, from very quiet environment to very loud, different places are selected for field testing of ASR systems. These include labs, offices, classrooms, campus-parking space, open-fields, cafeteria, bus-stand and roads within the campus. The demographics includes both technical and non-technical people working in the university and illiterate people like car drivers, rickshaw drivers, shopkeepers and waiters of the cafeteria. Best efforts are made to have an equal representation of males and females but in case of drivers, shopkeepers and waiters, it is not possible to get the system tested by the females. A total of 80 speakers have took part in the testing and each speaker spoke 7 to 8 different district names. The total test files are 586 excluding the files in which user didn't speak anything. Table III presents the results on the performance of standalone speech recognition systems.

TABLE III. Performance of ASR Systems in Field

ASR Type	Testing Utterances	Correctly Decoded	Accuracy (%age)
Accent Independent	586	441	75.25
Accent Dependent	586	352	60.06

It is quite clear that accent independent system clearly outperforms the accent dependent system. The reason for poor performance lies on the fact that accent of an unknown utterance is not identified with good accuracy. The isolated word accent-independent speech recognition system due to its better performance was integrated with weather information spoken dialog system and deployed at Pakistan Meteorological Department (PMD), Islamabad on a landline number. The initial version of the system was deployed on 13th August 2015 and the performance of the system was monitored for two months. The system gave an accuracy of 71% for 3560 utterances.

A. Improvements

The accuracy has dropped from 92% in laboratory to 71% in the field. In order to improve the accuracy of system in the field, the speech corpus was cleaned and verified again along with the addition of new recorded data. This modified system had an accuracy of 93.01% in laboratory testing. For 3095 utterances in the field, this system gave an accuracy of 79.46%. This improved accuracy is still quite lower than the accuracy in lab of 93%. In order to further improve the system, we decided to adapt the system to the users by including the field data in the training of the system. This field data means user responses recorded during the operation of the system in the field. The difference between field data and collected corpus is that the collected corpus is recorded under supervision which may not reflect exactly how the users will say their response in real world. As corpus is usually collected from some sections of population (mostly students in our case), it may not cover the target users effectively. The field data contains both of these aspects which can lead to improved accuracy of the system.

More than 7000 utterances recorded in the field were added to the training data. This resulted in a major turn around and helped in bringing the accuracy of the system closer to the laboratory results. The adapted ASR system had a field accuracy of 92.56% for 2609 utterances.

V. Conclusion

In this paper, we have presented Urdu speech recognition system for district names of Pakistan. To handle accent variation, accent dependent and accent independent ASRs have been developed and their tested. Accent dependent ASRs perform better when accent of the input is known and the ASR for that accent is used. However, the accent of the input is unknown in the field and accent identifier is used to determine the accent from the input. The performance of accent identifier is poor due to inadequate capturing of accent variation in isolated words that are proper nouns. This leads to the degradation in performance of accent dependent speech recognition system. The accent independent ASR system performs better than accent dependent system. However, further study is required to develop better accent identification system and then compare the accent independent and accent dependent systems. Furthermore, addition of field data in the training of speech recognition system helps to improve its accuracy in the field.

Acknowledgment

This work has been conducted through the project, Enabling Information Access for Mobile based Urdu Dialogue Systems and Screen Readers supported through a research grant from National ICT RnD Fund, Pakistan.

References

[1] K. H. Davis, R. Biddulph and S. Balashek, "Automatic Recognition of Spoken Digits," Journal of the Acoustical Society of America, vol. 24, no. 6, pp. 627-642, 1952.

[2] Y. Chow, M. Dunham, O. Kimball and M. Krasner, "BYBLOS: The BBN continuous speech recognition system," in International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Dallas, Texas, USA, 1987.

[3] V. Zue, J. Glass, M. Phillips and S. Seneff, "The MIT SUMMIT Speech Recognition system: a progress report," in HLT Workshop on Speech and Natural Language, Philadelphia, PA, USA, 1989.

[4] M. Adda-Decker, G. Adda, J. L. Gauvain and L. Lamel, "LARGE VOCABULARY SPEECH RECOGNITION IN FRENCH," in IEEE International Conference on Acoustics, Speech, and Signal Processing, Phoenix, AZ, USA, 1999.

[5] J. Sherwani, "Speech Interface for Information Access by Low-Literate Users in the Developing World," Pittsburgh, PA, USA, 2009.

[6] F. Qiao, J. Sherwani and R. Rosenfeld, "Small-vocabulary speech recognition for resource-scarce languages," in First ACM Symposium on Computing for Development, London, United Kingdom, 2010.

[7] M. A. M. A. Shariah, R. N. Ainon, R. Zainuddin and O. O. Khalifa, "Human computer interaction using isolated-words speech recognition technology," in International Conference on Intelligent and Advanced Systems (ICIAS), Kuala Lumpur, Malaysia, 2007.

[8] P. Saini, P. Kaur and M. Dua, "Hindi Automatic Speech Recognition Using HTK," International Journal of Engineering Trends and Technology (IJETT), vol. 4, no. 6, pp. 2223-2229, 2013.

[9] M. A. Hasnat, J. Mowla and M. Khan, "Isolated and Continuous Bangla Speech Recognition: Implementation, performance and application perspective," BRAC University, Dhaka, Bangladesh, 2007.

[10] J. Ashraf, N. Iqbal, N. S. Khattak and A. M. Zaidi, "Speaker Independent Urdu Speech Recognition," in International Conference on Informatics and Systems (INFOS), Cairo, Egypt, 2010.

[11] H. Sarfraz, S. Hussain, R. Bokhari, A. A. Raza, I. Ullah, Z. Sarfraz, S. Pervez, A. Mustafa, I. Javed and R. Parveen, "Large vocabulary continuous speech recognition for Urdu," in International Conference on Frontiers of Information Technology, Islamabad, Pakistan, 2010.

[12] K. F. Lee, H. L. Hon and R. Reddy, "An overview of the SPHINX speech recognition system," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 38, no. 1, pp. 35-45, 1990.

[13] S. Rauf, A. Hameed, T. Habib and S. Hussain, "District Names Speech Corpus for Pakistani Languages," in Oriental COCOSDA/CASLRE Conference, Shanghai, China, 2015.

[14] L. R. Rabiner and M. R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances," Bell System Technical Journal, vol. 54, no. 2, p. 297–315, 1975.

[15] Afsheen, S. Irtza, M. Farooq and S. Hussain, "Accent Classification among Punjabi, Urdu, Pashto, Saraiki and Sindhi," in Conference on Language and Technology (CLT), Karachi, Pakistan, 2014.