Chapter 2
Arabic Speech Recognition Systems

This chapter presents a brief overview of the evolution of Arabic speech recognition systems. It provides a literature survey of Arabic speech recognition systems and discusses some of the challenges of Arabic from the speech recognition point of view.

2.1 Literature and Recent Works

Development of an Arabic speech recognition is a multidiscipline effort, which requires integration of Arabic phonetic (Elshafei 1991; Alghamdi 2000; Algamdi 2003), Arabic speech processing techniques (Elshafei et al. 2002, 2007; Al-Ghamdi et al. 2003), and natural language processing (Elshafei et al. 2006). Development of an Arabic speech recognition system has recently been addressed by a number of researchers.

Recognition of Arabic continuous speech was addressed by Al-Otaibi (2001). He provided a single-speaker speech dataset for MSA. He also proposed a technique for labeling Arabic speech. He reported a recognition rate for speaker-dependent ASR of 93.78% using his technique. The ASR was built using the Hidden Markov Model (HMM) tool kit (HTK). Hyassat and Abu Zitar (2008) described an Arabic speech recognition system based on Sphinx4. They also proposed an automatic toolkit for building phonetic dictionaries for the Holy Qur’an and standard Arabic language. Three corporuses were developed in this work, namely, the Holy Qura’an corpus HQC-1 of about 18.5 h, the command and control corpus CAC-1 of about 1.5 h, and the Arabic digits corpus ADC of less than 1 h of speech.

A workshop was held in 2002 at John Hopkins University. Kirchhoff et al. (2003) proposed to use Romanization method for transcription of Egyptian dialectic of telephone conversations. Soltau et al. (2007) reported advancements in the IBM system for Arabic speech recognition as part of the continuous effort for the GALE project. The system consists of multiple stages that incorporate both vocalized and non-vocalized Arabic speech model. The system also incorporates a training corpus of 1,800 h of unsupervised Arabic speech. Azmi et al. (2008)
investigated using Arabic syllables for speaker-independent speech recognition system for Arabic spoken digits. The database used for both training and testing consists of 44 Egyptian speakers. In a clean environment, experiments show that the recognition rate obtained using syllables outperformed the rate obtained using monophones, triphones, and words by 2.68%, 1.19%, and 1.79%, respectively. Also in noisy telephone channel, syllables outperformed the rate obtained using monophones, triphones, and words by 2.09%, 1.5%, and 0.9%, respectively. Abdou et al. (2006) described a speech-enabled computer-aided pronunciation learning (CAPL) system. The system was developed for teaching Arabic pronunciations to non-native speakers. The system uses a speech recognizer to detect errors in user recitation. A phoneme duration classification algorithm is implemented to detect recitation errors related to phoneme durations. Performance evaluation using a dataset that includes 6.6% wrong speech segments showed that the system correctly identified the error in 62.4% of pronunciation errors, reported “Repeat Request” for 22.4% of the errors, and made false acceptance of 14.9% of total errors. Khasawneh et al. (2004) compared the polynomial classifier that was applied to isolated-word speaker-independent Arabic speech and dynamic time warping (DTW) recognizer. They concluded that the polynomial classifier produced better recognition performance and much faster testing response than the DTW recognizer. Choi et al. (2008) presented recent improvements to their English/Iraqi Arabic speech-to-speech translation system. The presented system-wide improvements include user interface (UI), dialog manager, ASR, and machine translation (MT) components. Rambow et al. (2006) addressed the problem of parsing transcribed spoken Arabic. They examined three different approaches: sentence transduction, treebank transduction, and grammar transduction. Overall, grammar transduction outperformed the other two approaches. Parsing can be used to check the speech recognizer n-best hypothesis to rescore them according to most syntactically accurate one. Nofal et al. (2004) demonstrated a design and implementation of stochastic-based new acoustic models suitable for use with a command and control system speech recognition system for the Arabic language. Park et al. (2009) explored the training and adaptation of multilayer perceptron (MLP) features in Arabic ASRs. Three schemes had been investigated. First, the use of MLP features to incorporate short-vowel information into the graphemc system. Second, a rapid training approach for use with the perceptual linear predictive (PLP) + MLP system was described. Finally, the use of linear input networks (LIN) adaptation as an alternative to the usual HMM-based linear adaptation was demonstrated. Shoaib et al. (2003) presented a novel approach to develop a robust Arabic speech recognition system based on a hybrid set of speech features. This hybrid set consists of intensity contours and formant frequencies. Imai et al. (1995) presented a new method for automatic generation of speaker-dependent phonological rules in order to decrease recognition errors caused by pronunciation variability dependent on speakers. Choueiter et al. (2006) concentrated our efforts on MSA, where they built morpheme-based LMs and studied their effect on the OOV rate as well as the word error rate (WER). Bourouba et al. (2006) presented a new HMM/support vectors machine (SVM) (k-nearest neighbor) for recognition of
isolated spoken words. Sagheer et al. (2005) presented a novel visual speech features representation system. They used it to comprise a complete lip-reading system. Taha et al. (2007) demonstrated a novel agent-based design for Arabic speech recognition. They defined the Arabic speech recognition as a Multi-agent System where each agent has a specific goal and deals with that goal only. Elmisery et al. (2003) implemented a pattern matching algorithm based on HMM using field programmable gate array (FPGA). The proposed approach was used for isolated Arabic word recognition and achieved accuracy comparable with the powerful classical recognition system. Mokhtar and El-Abddin (1996) represented the techniques and algorithms used to model the acoustic-phonetic structure of Arabic speech recognition using HMMs. Gales et al. (2007) described the development of a phonetic system for Arabic speech recognition. A number of issues involved with building these systems had been discussed, such as the pronunciation variation problem. Bahi and Sellami (2001) presented experiments performed to recognize isolated Arabic words. Their recognition system was based on a combination of the vector quantization technique at the acoustic level and Markovian modeling.


Alghamdi et al. (2009) developed an Arabic broadcast news transcription system. They used a corpus of 7.0 h for training and 0.5 h for testing. They achieved a WER of 8.61%. The WER obtained ranged from 14.9 to 25.1% for different types and sizes of test data. Satori et al. (2007) used Sphinx tools for Arabic speech recognition. They demonstrated the use of the tools for recognition of isolated Arabic digits. The data were recorded from six speakers. They achieved a digits recognition accuracy of 86.66%. Lamel et al. (2009) described the incremental improvements to a system for the automatic transcription of broadcast data in Arabic, highlighting techniques developed to deal with specificities (no diacritics, dialectal variants, and lexical variety) of the Arabic language. Afify et al. (2005) compared grapheme-based recognition system with explicitly modeling short vowels. They found that short vowels modeling improves recognition performance. Billa et al. (2002) described the development of audio indexing system for broadcast news in Arabic. Key issues addressed in this work revolve around the three major components of the audio indexing system: automatic speech recognition, speaker identification, and named entity identification.
Messaoudi et al. (2006) demonstrated that by building a very large vocalized vocabulary and by using a language model including a vocalized component, the WER could be significantly reduced. Elmahdy et al. (2009) used acoustic models trained with large MSA news broadcast speech corpus to work as multilingual or multi-accent models to decode colloquial Arabic. Vergyri et al. (2004) showed that the use of morphology-based language models at different stages in a large-vocabulary continuous speech recognition (LVCSR) system for Arabic leads to WER reductions. To deal with the huge lexical variety, Xiang et al. (2006) concentrated on the transcription of Arabic broadcast news by utilizing morphological decomposition in both acoustic and language modeling in their system. Selouani and Alotaibi (2011) presented Genetic Algorithms to adapt HMMs for non-native speech in a large-vocabulary speech recognition system of MSA. Saon et al. (2010) described the Arabic broadcast transcription system fielded by IBM in the GALE project. Key advances include improved discriminative training, the use of subspace Gaussian mixture models (SGMM), neural network acoustic features, variable frame rate decoding, training data partitioning experiments, unpruned n-gram language models, and NNLMs. These advances were instrumental in achieving a WER of 8.9% on the evaluation test set. Kuo et al. (2010) studied various syntactic and morphological context features incorporated in an NNLM for Arabic speech recognition.

2.2 Arabic Speech Recognition Challenges

Arabic speech recognition faces many challenges. For example, Arabic has short vowels which are usually ignored in text. Therefore, more confusion will be added to the ASR decoder. Additionally, Arabic has many dialects where words are pronounced differently. Elmahdy et al. (2009) summarized the main problems in Arabic speech recognition, which include Arabic phonetics, diacritization problem, grapheme-to-phoneme relation, and morphological complexity. Diacritization is represented by different possible diacritizations of a particular word. As modern Arabic is usually written in non-diacritized scripts, lots of ambiguities for pronunciations and meanings are introduced. Elmahdy et al. (2009) also showed that grapheme-to-phoneme relation is only true for diacritized Arabic script. Arabic morphological complexity is demonstrated by the large number of affixes (prefixes, infixes, and suffixes) that can be added to the three consonant radicals to form patterns. Farghaly and Shaalan (2009) provided a comprehensive study of Arabic language challenges and solutions. Lamel et al. (2009) presented a number of challenges for Arabic speech recognition such as no diacritics, dialectal variants, and very large lexical variety. Alotaibi et al. (2008) introduced foreign-accented Arabic speech as a challenging task in speech recognition. A workshop was held in 2002 at John Hopkins University to define and address the challenges in developing a speech recognition system using Egyptian dialectic Arabic for telephone conversations. They proposed to use Romanization method for transcription of the speech corpus (Kirchoff et al. 2003). Abushariah et al. (2010) reported the
design, implementation, and evaluation of a research work for developing a high-performance natural speaker-independent Arabic continuous speech recognition system. Muhammad et al. (2011) evaluated conventional ASR system for six different types of voice disorder patients speaking Arabic digits. MFCC and Gaussian mixture model (GMM)/HMM are used as features and classifier, respectively. Recognition result is analyzed for types of diseases. Billa et al. (2002) discussed a number of research issues for Arabic speech recognition, e.g., absence of short vowels in written text and the presence of compound words that are formed by the concatenation of certain conjunctions, prepositions, articles, and pronouns, as prefixes and suffixes to the word stem.

References


Cross-Word Modeling for Arabic Speech Recognition
AbuZeina, D.; Elshafei, M.
2012, X, 74 p. 15 illus., 1 illus. in color., Softcover
ISBN: 978-1-4614-1212-0