The fast pace of the advancement in information and communications technology is reshaping our society and vastly increasing our capabilities for faster learning, higher achievements, and better and wider communication, in addition to more effective and productive collaboration among speech scientists and engineers.

One of the important frontiers of communication technology is the user interface, namely, how can the man–machine interface be designed both in a more natural environment and a more immersive environment, which captures the essential attributes of a human-like exchange between human and machine. To address this important issue, researchers from various areas have been hard at work to equip machines with vital human-like capabilities, such as speech communication and vision. It is fair to say that despite many staggering technological successes achieved in these areas, the machine capabilities developed so far remain rather primitive compared to that of their human counterparts. This has propelled speech system designers to continue their relentless effort to achieve this far-reaching goal.

One such general area where research is continuing persistently is the speech processing area. Speech is the natural form of communication between humans. Its production is a highly nonlinear process that is strongly influenced by the high variability of factors such as age, gender, rate of speech, different dialects and regional accents, emotional state, and more. Speech perception is a hard task in that, in addition to the above-cited production-related difficulties, it has to contend with other equally variable and adverse factors such as background noise, interference from other speakers, room acoustics, recording equipment, and channel characteristics in the case of telephone conversation. Automatic speech recognition (ASR) is a key technology for a variety of applications, such as automatic translation, hands-free operation and control (as in cars and airplanes), automatic query answering, telephone communication with information systems, automatic dictation (speech-to-text transcription), and government information systems. In fact, speech communication with computers and household appliances is envisioned to be the
dominant human–machine interface in the near future. However, despite many impressive achievements in the area of speech recognition, reaching well-functioning human performance levels still remains a possibly unattainable goal.

During the last few decades, much research was carried out in the ASR area resulting in numerous practical and commercial successes with impressive high recognition performances, but only if the environment and the speaking manner are constrained such as with the use of isolated keywords.

No doubt, conversational or continuous speech recognition introduces many challenges to ASRs. One of these challenges is the pronunciation variation problem which known to reduce recognition accuracy. Pronunciation variation appears in the form of insertions, deletions, or substitutions of phoneme(s) relative to the canonical transcription of the words in the pronunciation dictionary. Within-word variations and cross-word variations (words’ junctures merging) are well-known variation problems in continuous speech. Accordingly, handling this phenomenon is a major requirement to have robust ASRs.

Within-words variation can be accounted for in ASR by using multiple pronunciation variants in the pronunciation dictionaries. However, cross-word variations alter the phonetic spelling of words beyond their listed forms in the pronunciation dictionary, leading to a number of out-of-vocabulary (OOV) word forms. This book presents a knowledge-based approach to model cross-word pronunciation variation at both pronunciation dictionary and language model levels. The proposed approach is based on modeling cross-word pronunciation variations by expanding the pronunciation dictionary and transcription corpus using modern standard Arabic (MSA) phonological rules.

The proposed method was tested using a baseline system that contains a pronunciation dictionary of 14,234 words from a 5.4-h pronunciation corpus of Arabic broadcast news. The expanded dictionary contains 16,273 words. Also, the corpus transcription is expanded according to the applied Arabic phonological rules. Using Carnegie Mellon University (CMU) Sphinx-III speech recognition engine, the enhanced system achieved a word error rate (WER) of 9.91% on a test set of fully diacritized transcription of about 1.1 h of MSA broadcast news. The WER is significantly reduced by 2.30% compared to the baseline system.

This book presents many examples in Arabic; a full appendix is provided for the Arabic terminologies used in this book.
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