Application of Speech Recognition Technology in Broadcast and Television Monitoring: Postprint

Authors: Zhao Yangyang

Date: 2023-10-08T00:00:00+00:00

Abstract

Nowadays, with the rapid development of science and technology, radio and television, as traditional media, have long served as a common channel for the public to receive news and information. Regarding the promotion of rapid progress in radio and television monitoring and supervision operations, the most efficient approach is to integrate speech recognition technology into radio and television monitoring and supervision, thereby effectively enhancing the quality and efficiency of such work. Based on this premise, to ensure the vigorous development of the radio and television monitoring and supervision industry, it is imperative to devote necessary attention and emphasis to the application and exploration of speech recognition technology within radio and television monitoring and supervision operations.

Full Text

Abstract

With rapid advancements in science and technology, radio and television—traditional media outlets—have long served as primary channels for news and information dissemination. Integrating speech recognition technology into broadcast monitoring and supervision operations represents the most effective approach to accelerating progress in this field, thereby significantly enhancing both quality and efficiency. To ensure the robust development of broadcast monitoring, focused attention must be devoted to the application and exploration of speech recognition technology.

Speech recognition technology, an interdisciplinary field drawing upon multiple domains of knowledge, has become widely deployed with ongoing technological progress. This technology converts input acoustic signals into corresponding commands or text, enabling task completion through voice alone without traditional input devices such as keyboards or mice. Since its inception over half

a century ago, speech recognition has evolved considerably. Early research focused on the Audry system, the first capable of recognizing several English letters. During the 1960s, advances in computer technology facilitated further development, with linear predictive analysis and dynamic programming emerging as key techniques for modeling speech signals. The 1970s brought major breakthroughs, including the maturation of dynamic time warping technology for speech alignment and improved vector quantization and Hidden Markov Model (HMM) theory. The 1980s saw deeper exploration and the proposal of various algorithms, most notably the combination of artificial neural networks with HMM models. The 1990s witnessed broader application, with numerous technology companies investing heavily in research and development. By the 21st century, broadcast news recognition rates reached approximately 90%, with current research focusing on multi-language simultaneous translation, spontaneous speech, and natural dialogue.

1. Speech Recognition Methods

Common speech recognition methods include: (1) **Linguistic and acoustic approaches**, which were among the earliest employed but remain limited by insufficient knowledge coverage, which has prevented their widespread adoption; (2) **Stochastic model methods**, which have achieved relatively mature application. This approach involves feature extraction, training module development, classification, and decision-making, incorporating HMM theory, Dynamic Time Warping (DTW), and Vector Quantization (VQ) techniques. Among these, the HMM algorithm stands out as the most straightforward and effective, offering superior recognition performance and thus being adopted by most systems; (3) **Neural network methods**, which emerged later in speech recognition development. These methods simulate human neural activity with capabilities such as autonomous learning and adaptation, demonstrating strong classification and mapping abilities. When combined with traditional approaches, they significantly improve recognition efficiency by leveraging respective strengths; (4) **Probabilistic grammar analysis**, a technique for recognizing long speech segments that distinguishes language features using multi-level knowledge to solve hierarchical problems, though it requires constructing appropriate and effective knowledge systems.

2. Speech Recognition Programs

Speech recognition programs primarily consist of four components: (1) **Working modes**, including command mode and recognition mode. Recognition mode operates with a background engine providing lexicon and recognition module libraries, requiring only modification of the main program source code without altering recognition grammar. Command mode, comparatively more complex to implement, requires programmers to compile dictionaries, perform coding, and conduct corrections based on speech lexicons. The key distinction is that programmers must modify and verify code according to dictionary contents; (2)

Environment configuration, typically including CTI server hardware parameter collection and setting, recognition hardware acquisition card initialization, and engine port configuration. All application operations rely on CTI technology. The platform determines whether speech input has occurred, acquires speech through the collection system, and utilizes speech cards for output and acquisition. In practice, boards within the speech card are activated by adding parameters to the program. Speech development platforms provide hardware API interface functions that operate through simple function calls and assignments; (3) **Speech dictionary compilation**, involving recognition rules, grammar, and speech template production according to platform standards. Dictionary setting requires configuring the core speech recognition package first, then implementing dictionary configuration based on self-compiled language standards; (4) **Main recognition program development**, the final compilation stage where programmers create a Graphical User Interface for user-computer interaction.

Numerous domestic and international vendors currently provide speech recognition technology. A comparative analysis of these platforms is presented in , revealing that each vendor' s technology possesses distinct advantages and characteristics, allowing enterprises to select appropriate solutions based on specific application scenarios.

Speech Recognition Technology	Platform Support	Features
Microsoft Speech API	Windows, iOS, Android, Web	High accuracy, audio transcription with timeline functionality
Google Speech API	iOS, Android, Web	Web online calling only
iFlytek Open Platform	iOS, Android, Windows, Linux	Supports Mandarin, Cantonese, and Sichuan dialect; accuracy exceeds 90%

Speech Recognition Technology	Platform Support	Features
	iOS, Android, Windows Phone, Web, Java, Linux	Supports Mandarin and Cantonese; accuracy above 95%, full-platform SDK

3. Application Categories in Broadcast Monitoring

3.1 Voiceprint Recognition

Voiceprint recognition technology analyzes speech waveforms to identify behavioral characteristics and determine speaker identity. It serves two purposes: evaluating speakers and verifying whether a particular voice matches a designated individual. Speech signals form the foundation of voiceprint recognition, enabling characterization of a speaker' s traits based on pronunciation patterns. As an important component of biometric authentication, voiceprint recognition shares similarities with fingerprint recognition in utilizing biological features, differing primarily in its focus on speaker-specific characteristics.

3.2 Content Identification

Content identification analyzes speech based on its physiological and physical properties to specifically evaluate and distinguish content, with the primary objective of determining the information carried by speech signals. However, significant room for improvement remains, as pronunciation habits and dialects directly impact recognition effectiveness. Voiceprint recognition can help address these challenges. Achieving consistency in grammar, semantics, and voiceprints requires comprehensive judgment through part-of-speech tagging, word differentiation, and contextual understanding, necessitating extensive comparative analysis within short timeframes.

3.3 Language and Speech Distinction

Language identification evaluates the linguistic characteristics compatible with speech materials, forming the basis for further research in content judgment and intelligent translation. This technology can assess multiple speech materials in computer systems, primarily through extraction by recognition systems. Additionally, comparison between standard speech models and individual speech patterns serves as the main method for identifying non-standard pronunciations during evaluation.

4. Practical Applications in Broadcast Monitoring

4.1 Application Scope

With continuous breakthroughs in speech recognition technology, automated systems now enable targeted assessment of real-time broadcast program status, extraction of key data, analysis of speech types and languages, and evaluation of speech signals, including silence and noise analysis. This facilitates simultaneous multi-spectrum research using speech recognition across channels. Integrating television content monitoring with speech recognition has substantially reduced human resource requirements while dramatically improving monitoring efficiency. Specific applications include:

(1) Television Monitoring: Flexible application enables the construction of speech and text templates for specific recognition tasks, real-time video recording, and accurate detection of broadcast anomalies. The monitoring system reports abnormalities to control stations with warning signals, allowing timely intervention to ensure safe broadcasting. Implementation of speech recognition technology can elevate system accuracy to 99%, achieving intelligent broadcast monitoring.

(2) Radio Monitoring: Language identification technology proves particularly crucial, as foreign radio stations broadcast in numerous languages at various times, requiring substantial manual effort for real-time monitoring. This challenge can be addressed by collecting and receiving speech recognition databases via satellite, then comparing recorded audio against library data to determine language types. However, significant physical data variations in speech present ongoing practical challenges. Introducing audio fingerprint similarity methods enables adaptive filtering through extensive learning, providing channel modeling capabilities.

4.2 Implementation Architecture

The system architecture comprises: **(1) Signal Demodulation Equipment**: Following source signal demodulation, the system invokes AM broadcast demodulators, cable TV demodulators, or FM broadcast demodulators as needed to convert broadcast signals into standard audio signals for recording by collection stations. Demodulator quantities can be selected based on monitoring channel requirements; **(2) Signal Preprocessing Equipment**: To maximize judgment effectiveness, AQC4 preprocessing equipment can be introduced to process audio signals, with control signal processors further correcting distortion to provide necessary source files for subsequent channel content monitoring; **(3) Multi-channel Collection Stations**: Cable TV demodulators extract video and audio to generate recognizable video and analog audio signals for recording by collection stations. These stations can receive and compress both broadcast and television signals for storage in server arrays, with separate stations for each signal type. Broadcast signal collection stations enable simultaneous recording of all broadcast signals with real-time volume display, soft mixing con-

sole integration for gain control, complete input signal monitoring, and scheduled recording designed for timeliness and efficiency, providing comprehensive listening services with automatic alerts for signal anomalies to prevent audio loss. Recording times can be adjusted according to broadcast schedules; **(4) Video Signal Collection Stations**: These stations can simultaneously acquire and record eight television channels, extracting audio signal codes from complete composite TV signals. Compression codes can be selected arbitrarily, with independent recording schedules designed for each channel based on broadcast times, typically using MPEG compression format. The interface displays all video images while enabling signal monitoring, with overall adjustment of saturation, chrominance, and contrast to ensure recording quality.

Conclusion

For broadcast monitoring personnel, radio frequency management and monitoring constitute critical responsibilities. With the popularization of radio technology and increasingly scarce frequency resources, broadcast monitoring is evolving toward full automation, necessitating the effective application of speech recognition technology to enhance monitoring quality and efficiency. To ensure robust development of broadcast monitoring, emphasis must be placed on effective, rational, and large-scale application of speech recognition technology.

References

[1] Zhang Jun. Application Analysis of Speech Recognition Technology in Content Supervision[J]. Television Guide, 2018(06): 254. [2] Li Zhiyuan. Overview of Speech Recognition Technology[J]. China New Telecommunications, 2018, 20(17). [3] Liu Gang, Liu Yang. Application Concept of Speech Recognition Technology in Radio Stations[J]. China Cable Television, 2015(11): 1300-1301. [4] Liu Lin, Sun Chengjiang, Wang Zhengxing. Exploration of Speech Recognition Technology Application in Intelligent Firefighting Construction[J]. China Management Informationization, 2018, 21(23): 174. [5] KJ.1029. Microsoft Computer Speech Recognition Technology Development Achieves Major Breakthrough[J]. Dual-Use Technology & Products, 2016(21): 20. [6] Sun Xiaojie. Implementation of Embedded Speech Recognition Technology[J]. Information Recording Materials, 2018, 19(08): 118-119. [7] Li Yue. Impact of Speech Recognition and Speech Synthesis Technologies on Broadcast Audio and Countermeasures[J]. Research on Transmission Competence, 2018, 2(05): 133-136. [8] Wang Haitao. Research on Audio and Speech Data Processing Technology in Radio and Television Monitoring Systems[D]. Northwestern Polytechnical University, 2007. [9] Guo Ligang, Fang Tufu. Application of Intelligent Sound Recognition Technology in Radio and Television Advertising Monitoring[J]. Radio & Television Technology, 2006(12): 72-74.

(Author affiliation: National Radio and Television Administration 281 Station)

Note: Figure translations are in progress. See original paper for figures.