Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/8/20230127

The development and status of speech recognition: An overview

Longwei Xiao

Beijing University of Posts and Telecommunications, 10 Xitucheng Road, Haidian, Beijing, PRC

2020213690@bupt.edu.cn

Abstract. Speech recognition has made remarkable progress in the last two decades, gradually moving from the laboratory to many different application scenarios. The core of speech recognition includes two processes: encoding and decoding, among which the decoding process can be divided into two parts: acoustic model and language model. After more than seventy years of development, speech recognition can be broadly divided into three phases in terms of technical direction: GMM-HMM era, DNN-HMM era, and end-to-end era. Finally, this paper will summarize the problems and future development direction of speech recognition through comparative study.

Keywords: speech recognition, DNN, HMM, deep learning.

1. Introduction

The core goal of speech recognition technology is to make machines understand human speech [1]. Digital signal processing, artificial intelligence, linguistics, mathematical statistics, acoustics, emotionology, and psychology are all included in this multidisciplinary intersection science. The most well-known products currently are Cortana from Microsoft and Siri from Apple. With the emergence of artificial intelligence, voice recognition technology has made significant advancements in both theory and practice and begun to move from the laboratory to the market [2].

After more than seventy years of development since 1952, there has been a unified consensus on the process of speech recognition. Speech recognition technology consists of two steps: encoding and decoding, while the decoding process can also be divided into four parts: input, acoustic model, language model and output. The development of speech recognition can be roughly divided into three stages [3]. The first stage is GMM-HMM, i.e., "HMM model using Gaussian mixture model to describe the probability distribution function of the vocal state", which was applied in 1970s; the second stage is DNN-HMM, i.e., DNN model is used to replace the GMM model in the previous stage [4]; the third stage is end-to-end, which is a one-step mapping of speech and text. In 2018, Google DeepMind proposed a large-scale CNN-RNN-CTC architecture that outperformed human experts by a factor of 6.

Speech recognition technology is still facing many problems, such as: (i) the problem of multiple speakers: how to distinguish between speakers; (ii) the complex acoustic environment: how to reduce the interference of noise on recognition; (iii) the problem of specialized vocabulary: how to recognize

DOI: 10.54254/2755-2721/8/20230127

specialized niche words; and (iv) the problem of dialects and multiple languages. A detailed discussion will be given in the following section [5].

2. The Process of speech recognition

Since 1952, speech recognition technology has been developed for more than 70 years and can be broadly divided into traditional recognition methods and end-to-end methods based on deep learning networks. Regardless of the method, it follows the process of "input-encoding-decoding-output", in which the decoding process can be divided into two parts: acoustic model and language model. As shown in Figure 1.

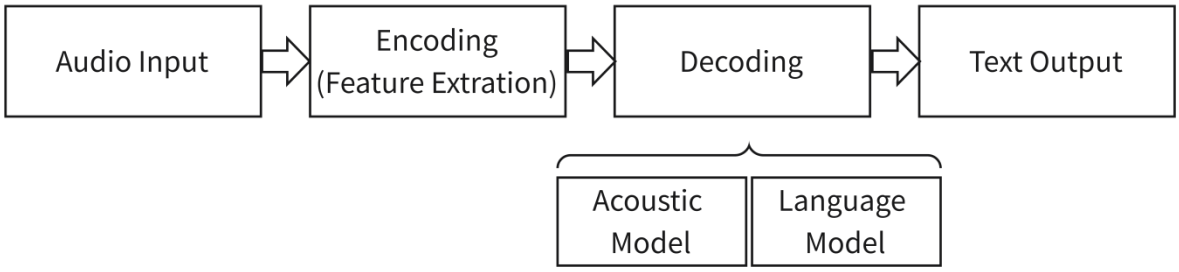

Figure 1. Speech recognition process.

2.1. Coding process

The input to speech recognition is sound, which is a signal that cannot be processed directly by a computer, so a coding process is needed to transform it into digital information and extract the features from it for processing. The encoding generally cuts the sound signal into small segments, called frames, at very short time intervals. For each frame, the features in the signal can be extracted by some rule (e.g., MFCC features) to turn it into a multidimensional vector. Each dimension in the vector is a feature of the signal of this frame, as shown in Figure 2 [6].

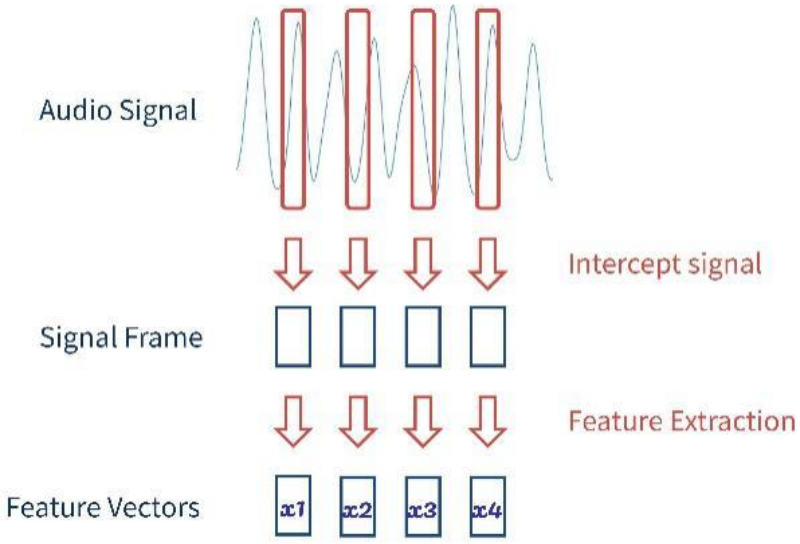

Figure 2. Speech recognition coding process.

2.2. Decoding process

The decoding process, on the other hand, is the process of turning the vectors obtained by encoding into words, which needs to be processed by two models, one model is the acoustic model, and the other is the language model [7]. The acoustic model processes the vectors obtained from the encoding

DOI: 10.54254/2755-2721/8/20230127

by combining adjacent frames into phonemes, i.e., vowels and rhymes in Chinese pinyin, and combining them into individual words or characters. The language model is used to adjust the illogical words obtained by the acoustic model to make the recognition result smooth [8]. To get an effective model, both require a large amount of data for training.

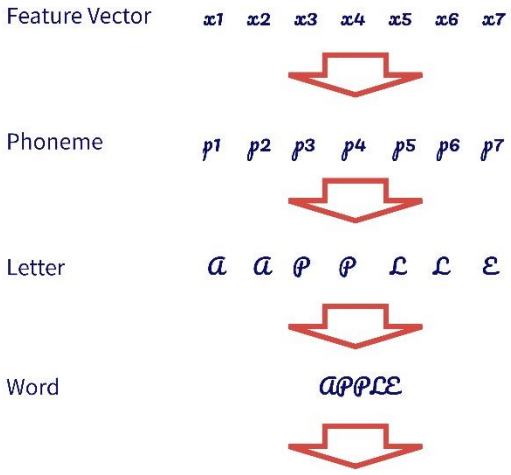

Figure 3. Language model processing.

An audio signal is known and processed into acoustic feature vectors represented as X= [x_1, x_2, x_3, ...], where xi denotes a frame feature vector; a possible text sequence is represented as W= [w_1, w_2, w_3, ...], where w_i denotes a word. The core of the speech model is to find the mapping relationship between W and X. The mathematical formula is expressed as follows.

$$W^* = \text{argmax}_w P(W|X)$$

This is the basic starting point of speech recognition. And it follows from the Bayesian formula that.

$$P(W \mid X) = \frac{P(X \mid W)P(W)}{P(X)} \propto P(X \mid W)P(W) \quad (1)$$

where P(X|W) is called the acoustic model and P(W) is called the language model.

2.3. Acoustic model

2.3.1. HMM and acoustic modeling. According to Equation 1 mentioned above, P(X|W) corresponds to the acoustic model, and the first thing to consider is that the indefinite length relationship between speech and text makes it impossible to have a one-to-one correspondence between the sequences of the two. the HMM model can be used to solve this problem [9].

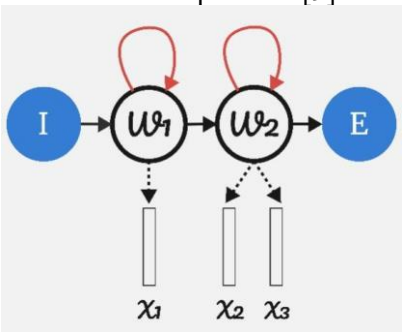

Figure 4. Hidden Markov chain model.(I, E for start and end)

DOI: 10.54254/2755-2721/8/20230127

Taking Figure 5 as an example, P(X|W)=P(x1,x2,x3|w1,w2) can be expressed in the form of the above HMM, where w is the implied state of the HMM and x is the observed value of the HMM, the number of implied states and the number of observations are not constrained by each other, which solves the problem of indefinite length of input and output, according to P(X|W)=P(x1,x2,x3|w1,w2), the following formula can be derived:

$$P(X \mid W) = P(w_1)P(x_1 \mid w_1)P(w_2 \mid w_1)P(x_2 \mid w_2)P(w_3 \mid w_2)P(x_2 \mid w_2) \quad (2)$$

Among them, the initial state probability $P(w_1)$ and the state transfer probability ($P(w_2|w_1)$, $P(w_3|w_2)$) of the HMM can be calculated from the total sample by conventional statistical methods, and the main difficulty lies in the calculation of the HMM emission probabilities ($P(x_1|w_1)$, $P(x_2|w_2)$ and $P(x_3|w_3)$), so the acoustic model problem is further refined to the learning of the HMM emission probabilities.

Another issue that needs to be addressed is the granularity size of the basic unit of text. For speech, the granularity of frames can be controlled by adjusting the width of the processing window. The word-level granularity is too broad and general, so, as shown in the figure, we decompose it into Phone, Triphone and Senone according to the granularity from wide to narrow.

Figure 5. Relationship between Phone, Triphone, and Senone.(#N, #N3, #3N3 indicate orders of magnitude.)

Each triphoneme can be modeled by an independent three-state HMM, so the basic units in terms of text are coded into tiny HMM states. Since many three-phonemes do not appear in the corpus or are small in number, and the states of the three-phonemes can eventually be shared through a decision tree, the number of three-phoneme states that are eventually retained for a language with a total of N phonemes is much less than $3N^3$, typically a few thousand, and they are generally defined as Senones, and the correspondence between each frame and each Senone is expressed as the three-phoneme HMM emission probability $P(x_i|s_j)$, where s_j denotes the jth Senone.

Sentence to Word, Word to Phone, Phone to Triphone, each Triphone is modeled by HMM, and a long chain of HMMs is formed by linking all related HMMs in order of pronunciation to represent Sentence.

2.3.2. GMM-HMM model. According to the above, the emission probability P(xi|sj) in the HMM of modeling will directly affect the acoustic model. Therefore, we can choose a Gaussian mixture model. Given a sufficient number of sub-Gaussians, the GMM can be fitted to an arbitrary probability distribution.

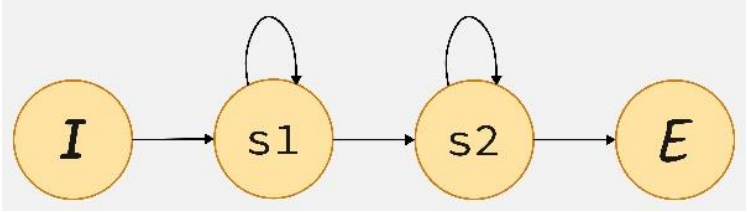

Figure 6. A three-phoneme GMM-HMM structure. (I, E for start and end.)

DOI: 10.54254/2755-2721/8/20230127

2.3.3. DNN-HMM model. DNN-HMM models need to be prepared with labels. Since the speech training set is often a correspondence between speech and whole text, the frame-level labels are not explicitly indicated. Therefore, additional algorithms are needed to label the datasets, and the method of choice is GMM as above. GMM is good at capturing the intrinsic relationships between known data, and the labeling has a high confidence level. The following figure shows the basic DNN-HMM acoustic model structure, where speech features are used as input to the DNN and the output of the DNN is used to calculate the emission probability of the HMM [10].

Compared to the GMM-HMM structure, the only difference between DNN-HMM and it is that the emission probabilities in the structure are derived from the DNN instead of the GMM [11]. Before training DNN-HMM, the target output value (i.e., label) of each speech frame on the DNN needs to be obtained. The labels can be obtained by training the Viterbi forced alignment of the GMM-HMM on the corpus. The DNN model is trained using the labels and input features [12], and the DNN model is used to compute the observation probabilities instead of the GMM, keeping other parts such as transfer probabilities and initial probabilities [11].

2.3.4. Language model. The problem to be solved by the language model is how to compute P(W), and the commonly used methods are based on n-element grammars or RNNs.

2.3.5. N Meta-Grammar Language Model. The language model is a typical Autoregressive Model, given a sequence W=[w1, w2,...wm], the probability is expressed as

$$P(W) = P(w_1, w_2, ..., w_m) = \prod_{i=1}^{m} P(w_i \mid w_1, w_2, ..., w_{i-1}) \propto \prod_{i=1}^{m} P(w_i \mid w_{i-n+1}, w_{i-n+2}, ..., w_{i-1}) \quad (3)$$

The above equation makes the assumption that the occurrence probability of the current word is only related to the n-1 words before the word, and each factor in the equation needs to be calculated statistically from a certain number of text corpus, and this process is the training process of the language model, and all possible $P(w_i \mid w_{i-n+1}, ..., w_{i-1})$ need to be listed.

The calculation process can be simplified by calculating the proportional relationship between the occurrence of the corresponding word strings in the corpus, i.e.,

$$P(wi \mid w_{i-n+1}, w_{i-n+2}, ..., w_{i-1}) = \frac{count(w_{i-n+1}, w_{i-n+2}, ..., w_i)}{count(w_{i-n+1}, w_{i-n+2}, ..., w_{i-1})} \quad (4)$$

where count indicates the number of occurrences of word strings in the corpus. Some word strings do not appear in the training text due to factors such as insufficient training corpus or uncommon word strings, which can be handled using the smoothing algorithm.

2.3.6. RNN language model. As can be seen by the sub-equation of the above probability formula, the current result depends on the previous information and therefore can be naturally modeled using a one-way recurrent neural network.

The conventional practice is to use the historical words in the sentence to predict the current word.

DOI: 10.54254/2755-2721/8/20230127

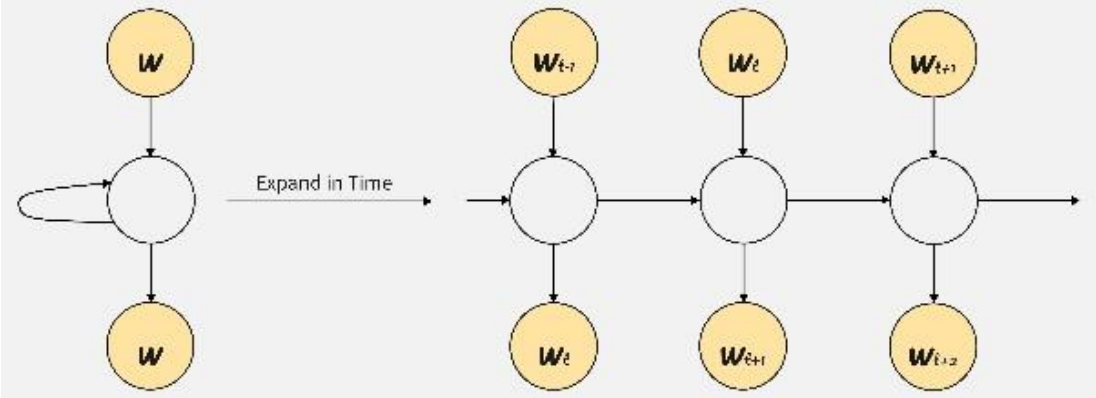

Figure 7. Basic structure of RNN language model.

2.3.7. End-to-End mode*l.* Due to the multivariate nature of speech and text, we started to consider the idea of end-to-end mapping of speech and text in one step to reflect end-to-end, but if the input is a whole speech and the output is the corresponding text, and both ends can be processed into a regular representation, we may be able to train a good end-to-end model as long as the data is sufficient and the model is appropriate [13].

2.4. CTC loss function

CTC is a loss function that has been applied to speech recognition since 2006. Given the training set, take one of the samples (X, W) as an example, input X into the model, the output can be any text sequence W', the probability of each text sequence is different, and we want the model to output W with as large a probability as possible, so the goal of CTC can be roughly understood as maximizing P(W|X) by adjusting the corresponding parameters of P.

The conventional structure in the use of CTC is LSTM-CTC, as shown in Figure 8 [14].

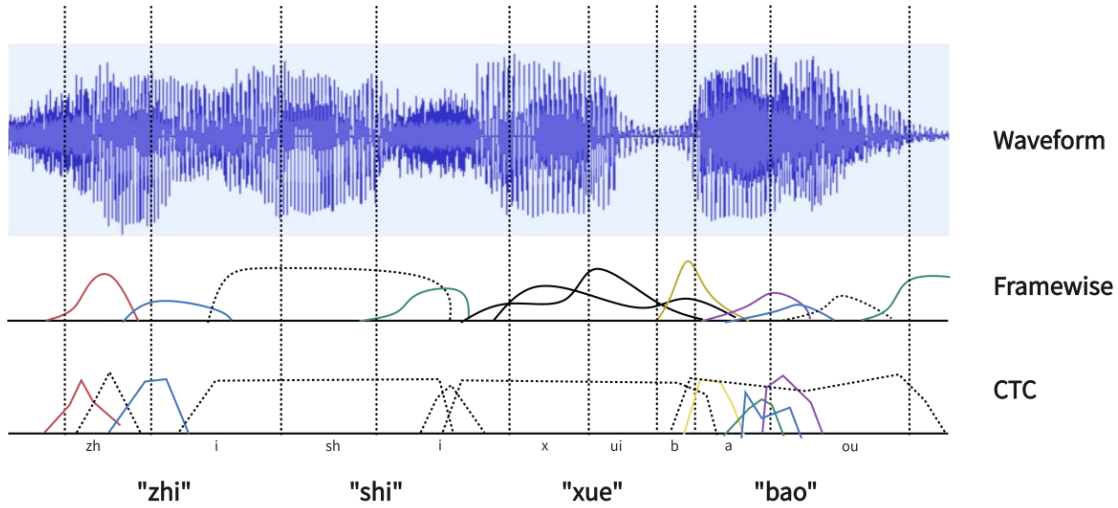

Figure 8. Comparison of prediction results between CTC sequence training and conventional frame-level training.

2.5. RNN-T technology

To achieve truly unified learning of acoustic and language models and improve system performance, RNN Transducer technology was proposed as early as 2012, and RNN-T technology was widely used until 2019 when Google successfully applied the technology to real-time offline speech recognition on mobile.

DOI: 10.54254/2755-2721/8/20230127

The idea of RNN-T explicitly training language models is to use the previous results as conditions when predicting the current target. The conditional probability p(y|x) is more intuitive in deep learning, where x is simply used as an input to predict y.

The following figure shows the structural differences between RNN-T and CTC.

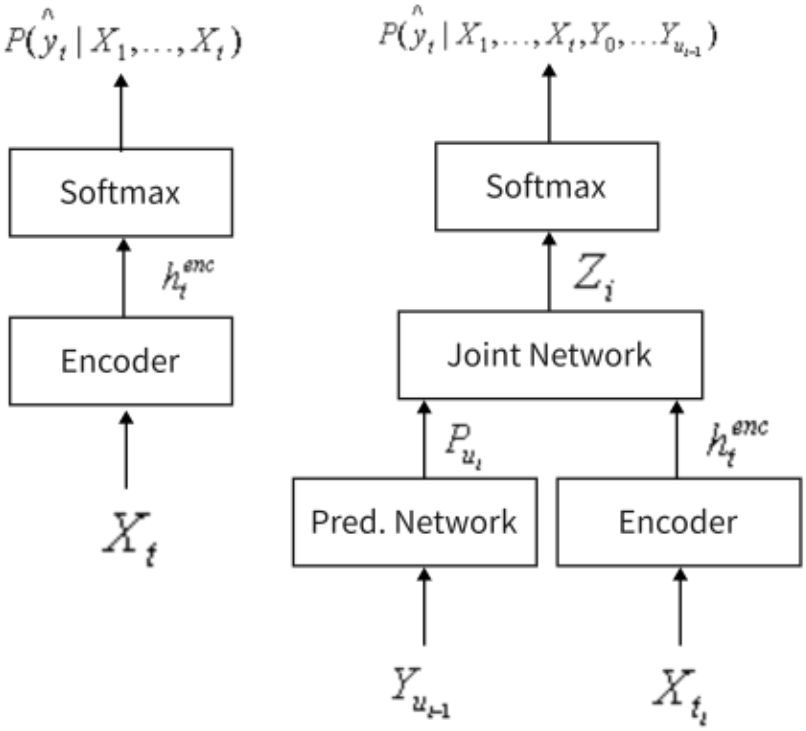

Figure 9. Structural differences between RNN-T and CTC.

where both the RNN-T prediction network and the encoder use LSTM, which can accumulate information on the historical output y and the historical speech features (x, with the current moment), respectively. and act on the new output jointly through a fully connected neural. p and h in the figure are the outputs of the prediction network and the encoder, respectively, in the form of a fixed-length vector.

3. History of speech recognition

Modern speech recognition can be traced back to 1952, when Davis et al. developed the world's first experimental system capable of recognizing the pronunciation of 10 English digits, which officially started the process of speech recognition.

Speech recognition can be roughly divided into three phases in terms of technology direction: GMM-HMM era, DNN-HMM era and end-to-end era.

DOI: 10.54254/2755-2721/8/20230127

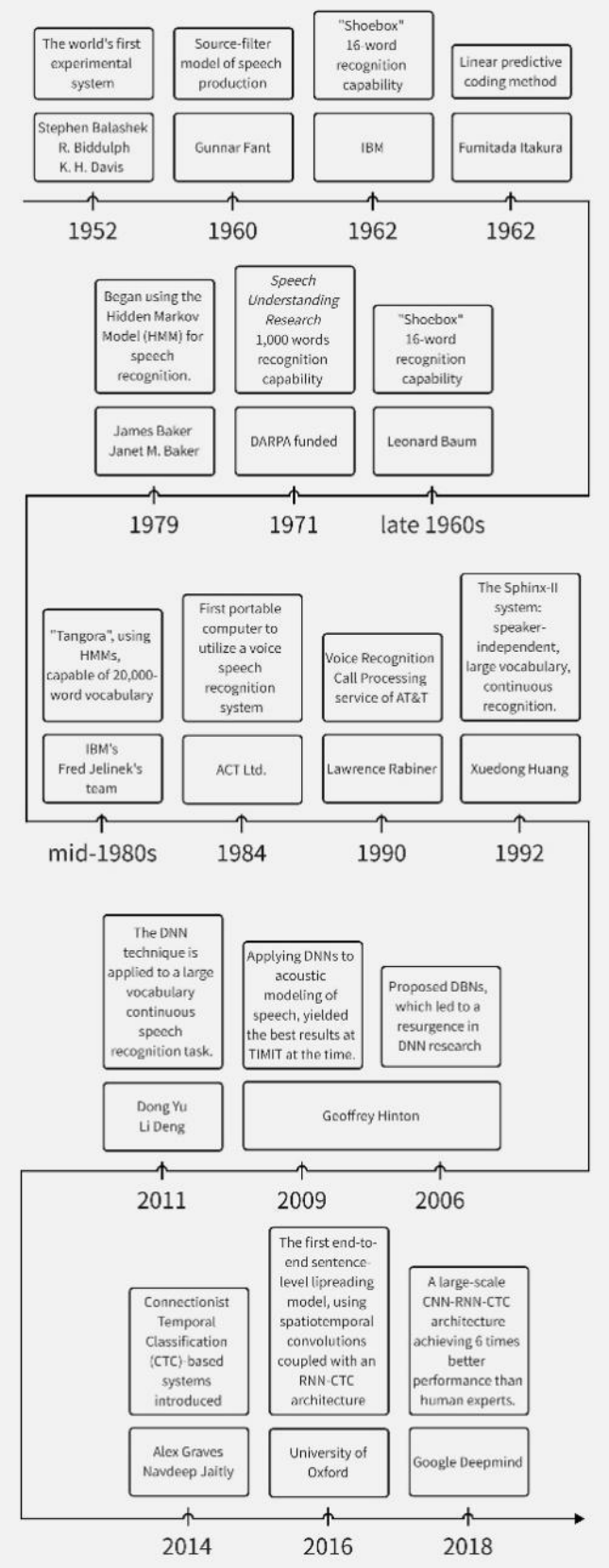

Figure 10. Stages of development of modern speech recognition technology.

DOI: 10.54254/2755-2721/8/20230127

3.1. GMM-HMM Era

3.1.1 Development of GMM-HMM model. The theoretical foundation of HMM was established around 1970 by Baum et al. and then applied to speech recognition by Baker at CMU and Jelinek at IBM.

GMM-HMM model, i.e., "HMM model using Gaussian mixture model (GMM) to describe the probability distribution function (PDF) of the vocal state" [15].

The physical meaning of HMM states can be considered as: the vocalization state of a phoneme, which is customarily divided into "initial state", "stable state" and "end state", and therefore can be modeled by three states the articulation of a phoneme. It is also possible to use two states to represent "start frame" and "other frames".

Researchers have proposed a variety of improvement techniques based on the GMM-HMM framework, including dynamic Bayesian methods with contextual information, discriminative training techniques, adaptive training techniques, and HMM/NN hybrid model techniques [16].

Since the discriminative training criterion and model adaption methods for acoustic models of speech recognition were proposed in the 1990s [17, 18], the progress of speech recognition has been rather gradual, and that line of speech recognition error rate has not greatly decreased.

3.1.2 Advantages of GMM-HMM model. GMM has a fast-training speed with a small acoustic model, and easy to port to embedded platforms.

3.1.3 Disadvantages of GMM-HMM GMM. doesn't use the contextual information of frames as well as cannot learn from deep nonlinear feature transformations (no activation function is used). The solution approach of GMM is based on EM algorithm, so it is possible to fall into local extremes.

To improve the recognition rate, the DNN model is used to replace the GMM model based on the three-phoneme HMM model, and a significant improvement in recognition rate is achieved [19].

3.2. DNN-HMM Era

3.2.1 Development of DNN-HMM model. Hinton's deep confidence network proposal in 2006 sparked a renaissance in deep neural network research. Hinton used DNNs in 2009 to predict the acoustics of speech, producing the greatest outcomes at TIMIT at the time. Late in 2011, Yu Dong and Deng Li of Microsoft Research introduced DNN approaches to a continuous voice recognition challenge with a huge vocabulary, significantly lowering the speech recognition error rate. Since then, the DNN-HMM age of voice recognition has begun.

In order to simulate each state, DNN-HMM mostly swaps out the original GMM model for a DNN model. The benefits of DNN include eliminating the need to make assumptions about the distribution of speech data, splicing adjacent speech frames, and containing the temporal structure information of speech, which significantly improves the classification probability for the state. DNN also has a powerful environment learning capability to increase the robustness to noise and accuracy [20].

3.2.2 Advantages of DNN-HMM model. The input features of DNN can be a fusion of multiple features, including discrete or continuous features, and DNN can make use of the structural information present in adjacent speech frames. The posterior probability distribution for estimating the state of an HMM using DNN does not require assumptions about the distribution of speech data.

3.2.3 Disadvantages of DNN-HMM model. Stitched frames allow for some degree of contextual information learning. The learned mapping connection is fixed from input to input due to the fixed window length of the DNN input, yet this results in a weaker modeling of the long-term temporal information correlation by the DNN.

DOI: 10.54254/2755-2721/8/20230127

3.3. End to End Era

3.3.1 Development of End-to-End model. The first end-to-end ASR effort was made in 2014 by Navdeep Jaitley of the University of Toronto and Alex Graves of Google DeepMind using a CTC-based system. Recursive neural networks and CTC layers make up the model. Using very huge datasets, Baidu expanded the work and found some commercial success in both Chinese Mandarin and English. LipNet [21], the first end-to-end sentence-level lip-reading model that employs spatiotemporal convolution and RNN-CTC architecture to achieve human-level performance in constrained grammar datasets, was proposed by Oxford University in 2016. A large-scale CNN-RNN-CTC architecture was proposed by Google DeepMind in 2018 and surpassed human experts by a factor of 6.

3.3.2 Advantages of End-to-End model. The end-to-end model uses a single objective function consistent with the ASR objective to optimize the entire network [22], while the traditional hybrid model optimizes each module individually and cannot guarantee global optimality. Since the end-to-end model uses a single network, which is more compact than the traditional hybrid model, the end-to-end model can be deployed to high-precision, low-latency devices.

3.3.3 Disadvantages of End-to-End model. The fact that end-to-end models learn language information differently from HMM-DNN models is a key factor in why these models do not outperform them. The language knowledge that the attention-based models learn is restricted within the transcriptional range of the training dataset, despite the fact that they are capable of learning language models. However the HMM-DNN model is distinct. It makes use of specialized dictionaries as well as a sizable language model that was trained using a sizable supplementary corpus of natural English. Its linguistic knowledge is highly extensive.

4. Existing problems with speech recognition

4.1. Scenarios and problems faced

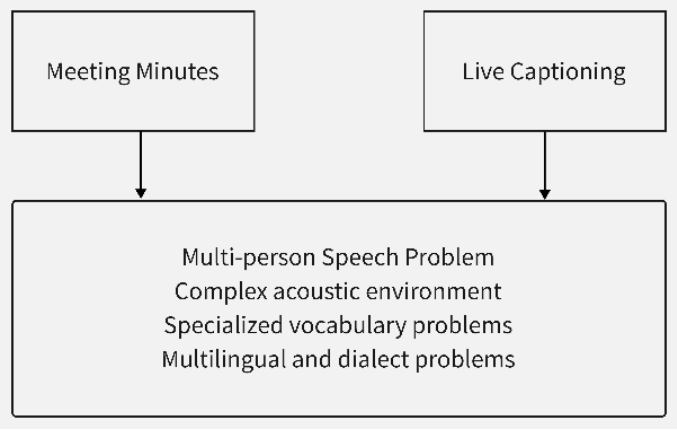

Figure 11. Possible application scenarios and problems.

Speech recognition systems are nowadays mainly used in voice assistant systems, such as Apple's Siri, Microsoft's Cortana and so on [23]. However, with the continuous development of technology, speech recognition systems nowadays face more complex scenarios and environments [24].

The application of speech recognition system in some other scenarios is still less. Take the two scenes of meeting recording and real-time captioning as an example, the reason for this is that the problems are mainly in the following four aspects.

DOI: 10.54254/2755-2721/8/20230127

a. Multi-person speech problem.
b. Complex acoustic environment problem
c. Professional vocabulary problem
d. Multilingualism and dialect problems

4.2. Multi-person speech problem

Traditional speaker classification systems are divided into two steps: the first step is to detect changes in the voice spectrum to determine when a speaker switch occurs; the second step is to identify each speaker in the conversation. This basic approach has many shortcomings, such as the fact that the change detection algorithm used here is not perfect and can result in segmented segments that may contain the speech of multiple speakers.

In a recent study, Google proposed a speaker classification system based on RNN-T, which reduced the speaker classification error rate from 20% to 2% and improved the performance by a factor of 10.

4.3. Complex acoustic environment problem

In the real environment, the speech signal is inevitably interfered by noise and reverberation, especially in the far-field conditions, as the sound waves in the process of propagation of its energy with the propagation distance is exponentially decayed, the speech signal is more serious interference by noise and reverberation, which greatly affects the performance of speech recognition and other speech interaction applications [25].

A approach for embedding noise robustness through contrastive learning into a method for contextually representing speech is proposed in the publication [26] as wav2vec-Switch. Even with a powerful language model for decoding, it manages to reduce relative word mistake rates by 2.9–4.9% on generated noisy Libri Speech data without degrading the original data and by 5.7% on real 1-channel noisy CHiME-4 data.

4.4. Professional vocabulary problem

In the practical application of speech recognition, the recognition effect is better for commonly used words, but there may be poor recognition accuracy for some unique names of people, songs, places, or proprietary words of a certain domain.

To solve the problem of specialized vocabulary recognition, a large amount of speech corpus containing specialized vocabulary can be collected, and then a domain-specific speech recognition model can be trained [27]. In practical applications, since it is difficult to collect speech corpus in specialized domains and the collected corpus is often insufficient, a general speech recognition model can be trained using a large amount of general corpus data first, and then do transfer learning training on the specialized domain corpus to get a speech recognition model suitable for the needs of specialized domains.

The method includes the following steps: Step 1. train a speech recognition model for professional domain by transfer learning; Step 2. collect professional vocabulary and update the finite-state transcriber; Step 3. update the finite-state transcriber of language model; Step 4. construct decoding space; Step 5. decode and recognize using HCLG files. This method can effectively solve the problem of low accuracy of speech recognition of specialized vocabulary in specific domains [28].

4.5. Multilingualism and dialect problem

The accent problem has always been a difficult problem inside speech recognition. To enhance the robustness of acoustic models to accent data, the technical solutions in the industry can be broadly classified into two types.

Multi-model scheme: according to the characteristics of accent pronunciation, the user is divided into multiple regions, each region corresponds to a model, and then the training data corresponding to

DOI: 10.54254/2755-2721/8/20230127

that region is used to train this regional model. The advantage of this scheme is that the model can be decoupled by region, but the model deployment and post-maintenance are relatively complicated.

Single-model scheme: According to the characteristics of accent pronunciation, users are divided into multiple regions, and then the region information is passed into the model as additional input, and this scheme finally uses only one model. This scheme ensures performance gains while simplifying the difficulty of model deployment and post-maintenance.

Although the multi-model scheme has better decoupling, the performance gain is relatively small, and there are more models to maintain and the model deployment is more troublesome; the single-model scheme is more effective, but less flexible and not convenient for targeted optimization of the product, for example, it is more difficult to do optimization for an individual accent.

For the multilingual speech recognition scenario, the Institute of Automation, Chinese Academy of Sciences has conducted research on multilingual speech recognition modeling methods. The main innovations are the proposed Shared-Hidden-Layer Multilingual LSTM (SHL-MLSTM) based on LSTM for multilingual speech recognition tasks with low resource data, and the proposed Multilingual ASR Transformer model, which eliminates the reliance on pronunciation dictionaries and the need for complex construction of universal tone subsets [29].

5. Conclusion

This paper first introduces the process of speech recognition, which is divided into two parts: encoding and decoding, where decoding can be divided into two parts: acoustic model and language model; then, this paper analyzes three development stages of speech recognition, and finally discusses the problems faced by four speech recognition systems based on two possible application scenarios and gives possible solutions. It is hoped that the exposition of this paper can bring readers thoughts on speech recognition.

References

[1] Ar E. Turkish Dictation System for Radiology and Broadcast News Applications.

[2] Future of eCommerce Development - 10 Trends Not to Miss in 2020. https://www.unifiedinfotech.net/blog/e-commerce-web-development-design-trends-for-2017/

[3] Proceedings. 9th IEEE International Workshop on Cellular Neural Networks and their Applications (IEEE Cat. No. 05TH8814) [C]// 2005 9th International Workshop on Cellular Neural Networks and Their Applications.

[4] Fan, Ruchao. 2018 [IEEE 2018 International Conference on Audio, Language and Image Processing (ICALIP) - Shanghai, China (2018.7.16-2018.7.17)] 2018 International Conference on Audio, Language and Image Processing (ICALIP) - CNN-Based Audio Front End Processing on Speech Recognition [J]. pp123-127.

[5] Jing Liu, Xinguang Xiang. [IEEE 2017 12th IEEE Conference on Industrial Electronics and Applications (ICIEA) - Siem Reap, Cambodia (2017.6.18-2017.6.20)] 2017 12th IEEE Conference on Industrial Electronics and Applications (ICIEA) - Review of the Anti-Noise Method in the Speech Recognition Technology [C]// 2017 12th IEEE Conference on Industrial Electronics and Applications (ICIEA).

[6] Florez-Choque O, Cuadros-Vargas E. 2007 Improving Human Computer Interaction through Spoken Natural Language[C]// IEEE Symposium on Computational Intelligence in Image & Signal Processing. IEEE.

[7] Zhu H, Deng Y, Ding P, et al. Apparatus and Method for Training A Neutral Network Acoustic Model, And Speech Recognition Apparatus and Method.

[8] Handling OOVWords in Mandarin Spoken Term Detection with an Hierarchical n-Gram Language Model[J]. Chinese Journal of Electronics, 2017(06):1239-1244.

[9] Candy J V. Discrete Hidden Markov Model Bayesian Processors[C]// Bayesian Signal Processing: Classical, Modern, and Particle Filtering Methods.

DOI: 10.54254/2755-2721/8/20230127

[10] Andrew H, Cheryl H, Louise H, et al. 2016 Does smoke-free Ireland have more smoking inside the home and less in pubs than the United Kingdom? Findings from the international tobacco control policy evaluation project[J]. European journal of public health, Vol. 18, No. 1, 2008:63-65.
[11] Li L, Yong Z, Jiang D, et al. 2013 Hybrid Deep Neural Network--Hidden Markov Model (DNN-HMM) Based Speech Emotion Recognition[C]// Affective Computing & Intelligent Interaction. IEEE.
[12] Wang S, Clark R, Wen H, et al. 2018 End-to-end, sequence-to-sequence probabilistic visual odometry through deep neural networks[J]. The International journal of robotics research, 37(4-5):513-542.
[13] Gan, Z.Y., Jia, H.J., Ruan, W.B., et al. 2016 Chinese-Tibetan cross-language voice conversion method and system,.
[14] Franyell Silfa, Jose-Maria Arnau, Antonio Gonzàlez et al. Boosting LSTM Performance Through Dynamic Precision Selection, Computer Architecture Deparment, Universitat Politecnica de Catalunya.
[15] Shiu, Yu; Kuo, C. -C. J. et al. 2005 Music genre classification via likelihood fusion from multiple feature models, Proceedings of the SPIE, Volume 5682, p. 258-268.
[16] Abdel-Hamid O, Hui J. 2013 Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code[C]// IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2013. IEEE.
[17] Saon G. 2006 A Non-Linear Speaker Adaptation Technique using Kernel Ridge Regression[C]// Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on. IEEE.
[18] Jayamaha R, Senadheera M, Gamage T, et al. 2009 VoizLock - Human Voice Authentication System using Hidden Markov Model[C]// International Conference on Information & Automation for Sustainability. IEEE.
[19] Fan R, Liu G. 2018 CNN-Based Audio Front End Processing on Speech Recognition[C]// 349-354.
[20] Meng X, Liu C, Zhang Z, et al. 2014 Noisy training for deep neural networks[C]// 2014 IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP). IEEE.
[21] Assael Y M, Shillingford B, Whiteson S, et al. LipNet: End-to-End Sentence-level Lipreading[J]. 2016.
[22] Ochiai T, Watanabe S, Hori T, et al. 2017 A Unified Architecture for Multichannel End-to-End Speech Recognition with Neural Beamforming[J]. IEEE Journal of Selected Topics in Signal Processing, (8):1-1.
[23] “Voice Search - Are you ready for the voice revolution?” https://www.innovationvisual.com/services/organic-search-seo/voice-search
[24] “HOW DOES SPEECH RECOGNITION WORK?” https://master-artificialintelligence.com/how-speech-recognition-work/
[25] elson, richard n., moutz, mitchell dubu, houte, james k. Acoustic evaluation and / or control of the fluid contents of the reservoir, JP4559218B2[P]. 2010.
[26] Wang Y, Li J, Wang H, et al. 2021 Wav2vec-Switch: Contrastive Learning from Original-noisy Speech Pairs for Robust Speech Recognition[J].
[27] Shinoda K. 2005 Speaker adaptation techniques for speech recognition using probabilistic models[J]. Electronics and Communications in Japan (Part III Fundamental Electronic Science), 88(12):25-42.
[28] Kim C W, Eom K W, Lee J W, et al. Signal Separation System and Method for Automatically Selecting Threshold to Separate Sound Sources: Us, Us20110182437 A1[P].
[29] Shiyu Zhou, 2018 Research on Multilingual Speech Recognition for Low-resource Languages, Beijing, Chinese Academy of Sciences (CAS).