***Previously on...***

Earlier I focused more on the speech parameters, research questions and results extracted from reduced speech research (addressing the "what to do" question), but after the Christmas holiday I wanted to get a clearer picture of how phonetic reduction research has a specific impact on improving the performance of automatic speech recognition systems (addressing the "why to do it" question), so I read some articles on ASR technology and took a few related online classes. Through the literature, I have gained a deeper understanding of the significance of the reduction research. Below is a review of recent reading on ASR. Although my final thesis is still more of a linguistic perspective on the acoustic features of reduced speech, it is necessary to understand the value and application of this perspective in the computer science field (especially the construction of pronunciation dictionary), and perhaps this will also be reflected in the feasibility of my subsequent proposal design and the long-term maintenance of my research interests.

**Methodological issues of ASR on using corpus**

Voice is one of the most economical and direct ways of interpersonal communication. With the development of information technology, it is expected that smart devices will have the ability to communicate with people. For this, intelligent devices must be able to recognize and *understand* speech signals. Automatic Speech Recognition (ASR) is the process of converting human speech signals into text or instructions (GB/T21023-2007)[[1]](#footnote-1). Currently, it has been widely used, such as voice dialing, voice control and voice query, and voice assistant for cell phone clients.

Automatic speech recognition has been studied for more than sixty years[[2]](#footnote-2). In 1952, K.H. Davis et al. of AT&T Bell Laboratory implemented Audry[[3]](#footnote-3), the first speech recognition system for isolated alphanumeric digits of a specific speaker, which targets the vowel part of isolated alphanumeric digits and extracts their resonance peak trajectory information for speaker-specific isolated digit recognition using a simple template matching method. Research in this period focused on the recognition of vowels, consonants, and simple isolated words[[4]](#footnote-4) [[5]](#footnote-5), with the hope of exploring more accurate acoustic features to describe phonological phenomena. In 1959, P. Denes et al. at the University of London constructed a phoneme recognition system that could recognize four vowels and nine consonants, which was the first to use statistical principles to improve phoneme recognition of multiphonemic words using statistical information about the sequence of phonemes that could be used as a qualification[[6]](#footnote-6) [[7]](#footnote-7). And in the same year, the Lincoln Laboratory at MIT implemented a recognition system for 10 vowel units, which, although still using resonant peak information, recognized no longer speaker-specific, but for non-speaker-specific[[8]](#footnote-8).

In the 1960s, Linear Predictive Coding (LPC) and Dynamic Time Warping (DTW) techniques were proposed[[9]](#footnote-9), which became the two major technological breakthroughs of this period. Although research in this period was still dominated by Isolated Word Recognition (IWR), in the mid-1960s, D.R. Reddy at Carnegie Mellon University (CMU) attempted to develop a continuous speech recognition system by dynamically tracking phonemes, and this work started the pioneering research on continuous speech recognition[[10]](#footnote-10).

In the 1970s, the field of speech recognition was still mainly concerned with isolated word recognition, and its breakthrough development was the introduction of the idea of pattern recognition into speech recognition by Velichko and Zagoruyko in Russia[[11]](#footnote-11). A few research institutions, represented by IBM and Bell Labs, began to shift their research from isolated word recognition to Large Vocabulary Continuous Speech Recognition (LVCSR). During this period, the Advanced Research Projects Agency (ARPA) of the U.S. Department of Defense organized a Speech Understanding Research (SUR) program for five institutions, including Carnegie Mellon University. During the five years of the program, speech recognition research received a huge boost, while more research institutions joined the speech recognition research team[[12]](#footnote-12) [[13]](#footnote-13). Under the support of the program, many excellent speech recognition systems and speech understanding systems were developed, such as, IBM's dictation-oriented machine[[14]](#footnote-14), BBN's HWIM system[[15]](#footnote-15), CMU's HEARSAY system[[16]](#footnote-16) and HAPPY system[[17]](#footnote-17), and Bell Labs' speech recognition system for telecommunication services.

In the 1980s, Connected Word Recognition (CVR) was also developed to recognize consecutive words by pattern matching of individual words based on isolated words. For connected word recognition, algorithms such as the two-layer dynamic programming method[[18]](#footnote-18), the JSRU over-all decoding method[[19]](#footnote-19), the hierarchical construction time regularization algorithm[[20]](#footnote-20), and the time-synchronized network search[[21]](#footnote-21) were proposed one after another. And the most significant breakthrough in speech recognition during this period was the shift from pattern matching approach to statistical model-based approach. With the successful application of Hidden Markov Model (HMM) to large vocabulary continuous speech recognition by IBM and Bell Labs[[22]](#footnote-22) [[23]](#footnote-23), various research units also gradually adopted the HMM-based speech recognition research line. In the mid-late 1980s, Artificial Neural Network (ANN) was reapplied to speech recognition[[24]](#footnote-24) [[25]](#footnote-25) [[26]](#footnote-26) [[27]](#footnote-27). Later, along with the emergence of statistical language models, HMM and language models jointly pushed the speech recognition technology to a new peak. Currently, most speech recognition systems in the world employ both techniques. During this period, several well-known speech recognition systems emerged, such as the world's first HMM-based speaker-specific, large-vocabulary continuous speech recognition system SPHINX[[28]](#footnote-28) and BNN's BYBLOS system[[29]](#footnote-29), which was successfully implemented by Kai-Fu Lee at CMU.

After the 1990s, the focus of speech recognition shifted to continuous speech, and the design and refinement of acoustic model structures, parameter extraction and optimization, and adaptive techniques led to further refinement of speech recognition systems and began to experiment with them in some specific domains. At this stage, both the theory and practice concerning the application of HMM in speech recognition have been further refined[[30]](#footnote-30) [[31]](#footnote-31) [[32]](#footnote-32), and some other key techniques have emerged, such as decision tree state clustering techniques for acoustic model parameter binding, and HMM-based acoustic model adaption techniques. As for language models, there has been a shift from rule-based approaches to statistical-based approaches, such as N-metric grammatical language models, resemble language models[[33]](#footnote-33) [[34]](#footnote-34), variable-length language models[[35]](#footnote-35) [[36]](#footnote-36) [[37]](#footnote-37), maximum entropy language models[[38]](#footnote-38) [[39]](#footnote-39) [[40]](#footnote-40) , and cache-based language models[[41]](#footnote-41). During this period, many research institutions or companies launched open-source or semi-open-source speech recognition systems, which set off a boom in the productization of speech recognition technology, such as the Via-voice system from IBM, the DRAGON system, and the HTK (Hidden Markov Tool Kit, HTK) system from the University of Cambridge, UK[[42]](#footnote-42), etc. In the late 1990s, speech from 1997 to 2001, the National Institute of Standards and Technology (NIST) organized an international evaluation of speech recognition for telephone conversations. evaluation. Compared with broadcast news speech, telephone conversations have more severe channel noise, more spoken words and dialects, and more speech overlap between speakers, which makes speech recognition more difficult and poses a great challenge. Many researchers have proposed many solutions to improve the performance of speech recognition by addressing the pronunciation variation and confusion in spoken and conversational speech.

After entering the 21st century, due to the rapid development and popularity of Internet and mobile Internet technologies, the development of computing power and network speed of smart phones have provided a new platform for the development and application of speech recognition technology. And the development of computer hardware technology and the emergence of cloud computing technology have also greatly promoted the research and application of speech recognition. First, the application mode of speech recognition has changed from "stand-alone mode" of cell phone client to "cloud mode", that is, the application of speech recognition is no longer limited by the hardware conditions of cell phone or computer terminal, but only the front-end is retained on the terminal device, and the recognition decoding is placed to the cloud server. In this way, not only the service provider can update and optimize the model at any time to provide better service, but also the user terminal does not need to save the model and calculation, which reduces the requirement for terminal hardware equipment and ensures the service quality at the same time. With the continuous popularity and rapid development of mobile devices, a series of speech recognition applications have emerged, among which Apple's Siri voice assistant for iPhone and Google's voice search system are the most famous. Second, thanks to the development of computer hardware, the speech recognition decoder based on Weighted Finite-State Transducer (WFST) has been further improved theoretically[[43]](#footnote-43) and has been more widely used. In addition, the modeling techniques for acoustic models, the adaptive techniques for acoustic models, and the discriminative training of acoustic models have also advanced greatly. In recent years, the training of acoustic models based on context-dependent Deep Belief Nets (DBNs) has achieved great success[[44]](#footnote-44). The architecture of large vocabulary continuous speech recognition has gradually changed from the original GMM-HMM to DNN-HMM. during this period, there has also been a breakthrough in language modeling techniques involving large-scale distributed language model training[[45]](#footnote-45) [[46]](#footnote-46), discriminative language model modeling[[47]](#footnote-47) [[48]](#footnote-48) [[49]](#footnote-49) [[50]](#footnote-50), neural network language modeling[[51]](#footnote-51) [[52]](#footnote-52) and recurrent neural network language modeling[[53]](#footnote-53) [[54]](#footnote-54), etc. And for different languages, both acoustic models and language models have been improved accordingly considering their characteristics. For example, for the speech recognition of adherent languages, such as Turkish, the modeling units of the acoustic model are mostly chosen as morphemes, and the composition of words is considered in the language modeling, and words are cut into roots and affixes to reduce the data sparsity in the modeling. For Chinese speech recognition, considering that Chinese is a phonetic language, the basic unit of writing is word, and each word has corresponding syllable representation and various ways of word composition in Chinese, the modeling unit of acoustic model can consider phonemes, phonemes and syllables, while the word items in language model and pronunciation dictionary can be selected with comprehensive consideration of word frequency, word length, sparsity problem and confusion.

Speech recognition system framework

The goal of large vocabulary continuous speech recognition is to find the sequence that optimally matches the input speech signal to it. The input speech signal, through front-end processing, is converted into a sequence of acoustically observed feature vectors, and then a sequence of optimally matched words is searched for. The language model and acoustic model provide the most important sources of knowledge for speech recognition systems. The basic process of the mainstream automatic speech recognition system based on Hidden Markov Model includes several parts, including feature extraction, acoustic model, pronunciation dictionary, language model and decoder. Here are basic instructions of these modules.

Feature Extraction

Feature Extraction is the first part of speech recognition, which extracts the observed feature vector sequence from the input speech signal that can be acoustically modeled. In systems based on Gaussian Mixture Model (GMM) for acoustic modeling, the Mel-Frequency Cepstral Coefficients (MFCC)[[55]](#footnote-55) and Perceptual Linear Predictive (PLP)[[56]](#footnote-56) are two commonly used speech features, and their first order and second-order differences are also often used to characterize the correlation between speech frames. In systems based on Deep Neural Network (DNN) for acoustic modeling, the Filter Bank features[[57]](#footnote-57), where multiple frames are stitched together, and the TRAPs (Temporal Patterns) features[[58]](#footnote-58), which are based on the banding energy, are more commonly used. Usually, certain front-end signal processing techniques are also adopted at this stage for reducing the effects of ambient noise, channel, speaker, etc.

Acoustic Model

The purpose of the Acoustic Model (AM) is to provide a method that, for a given sequence of acoustically observed features of a word, it is possible to compute its likelihood. In principle, the required probability distribution can be obtained by finding many samples of a given word. However, for large vocabulary continuous speech recognition tasks, this is not realistic. First, new words always appear without corresponding samples; second, there are so many words and different words have different acoustic realizations that it is impossible to exhaust all samples. Therefore, the word sequence is split into subword sequences, such as phonemes, the basic unit of speech. In this way, this module is more closely related to phoneme modeling.

Hidden Markov model is the current mainstream modeling approach for acoustic modeling. It assumes that the human articulatory process can be portrayed by an implicit sequence of states, where each state outputs a sequence of acoustically observed features with a certain probability distribution. In 1989, the excellent work of Rabinar et al. laid the theoretical foundation for the successful application of HMM in speech recognition[[59]](#footnote-59).Three basic problems of HMM were solved.

(1) Evaluation problem: The parameters of the HMM and the observed feature vector sequence are known, and the likelihood of the model to produce this observed feature vector sequence can be calculated using the forward algorithm.

(2) Decoding problem: The parameters of the HMM and the sequence of observed feature vectors are known, and the Viterbi decoding algorithm based on the idea of dynamic programming is used to search for the implicit state sequence in the HMM that is most likely to generate the sequence of observed feature vectors.

(3) Learning problem: Based on the known training data, the Baum-Welch algorithm[[60]](#footnote-60) based on the Maximum Likelihood Estimation (MLE) criterion and the Expectation Maximization (EM) idea[[61]](#footnote-61) [[62]](#footnote-62)can effectively learn the parameters of the model.

HMM is a powerful support for acoustic modeling. It provides a powerful way to integrate segmentation, temporal regularization, pattern matching, and contextual knowledge in a unified way. Most of the speech recognition systems use phonemes as the modeling unit of acoustic models, i.e., each phoneme is represented by an HMM. And due to the temporal nature of speech, the models generally adopt a left-to-right topology that allows for self-skipping and cross-state jumping. To address the effect of co-articulation, phoneme modeling is usually extended to context-dependent three-phoneme modeling, but this brings another side effect of model number inflation. To address this problem, a commonly adopted strategy is to introduce decision trees to cluster states with similar output probability distributions, thus reducing the requirement for the amount of training data.

Language Model

Language Model is one of the important components of automatic speech recognition system, which provides the probability of word sequences in the decoding process, and its essence is to provide the decoder with the word search path information from the higher linguistic layer to limit the decoding path, thus improving the decoding speed and judging more accurate recognition results. It encodes syntax, semantics, and pragmatics simultaneously, and focuses only on local dependencies. In addition, the probability distribution of N meta-grammar can be calculated directly from text data without the need of explicit linguistic rules. However, the N meta-grammar language model faces the problem of data sparsity because the training data is limited, and it is difficult to cover the whole linguistic phenomena. In addition, due to the characteristics of the language itself, the collocation use of certain words is rarely used in natural language or real life, and even if the training data is expanded, the distribution of N meta-grammar is still unbalanced, and the sparsity problem still exists. Therefore, it can be said that the inherent problem of data-sparse N meta-grammar cannot be overcome fundamentally but can only be alleviated to some extent by other techniques.

Pronunciation Dictionary

The Pronunciation Dictionary (PD) is one of the important components in continuous speech recognition. Speech recognition can only recognize the entries contained in the dictionary. In addition, the pronunciation dictionary contains the set of pronunciation units in speech recognition, which realizes the mapping between the modeling units of the acoustic model and the modeling units of the language model. It has an influence on the language model on the one hand, and a constraint on the acoustic model on the other. The decoder construction process can know which pronunciation units each word consists of according to the pronunciation lexicon, so that the sequence of acoustic modeling units corresponding to the word can be obtained.

Pronunciation dictionaries have an important role and impact on continuous speech recognition.

First, speech recognition can only recognize the entries contained in the pronunciation lexicon.

Second, the basic units of acoustic modeling in large-vocabulary continuous speech recognition are subwords, including syllables or morphemes (stems + affixes of adherent words), etc. In other words, the acoustic modeling units of large vocabulary continuous speech recognition are closer to the physiological structure of pronunciation and are disconnected from the linguistic content to be expressed. And the higher up to the linguistic level the more the content of the sentence is expressed. Pronunciation is a bridge between the acoustic layer (physiological layer) and the linguistic layer, which couples the content of the acoustic and linguistic layers.

Third, confusion is an inherent phenomenon in speech that is unavoidable, and the degree of confusion in speech is exacerbated by various articulatory variations in continuous speech, which leads to confusion within the acoustic model in automatic speech recognition. Continuous speech perception experiments have shown that acoustic information alone is not sufficient to correctly recognize speech; knowledge of the context of speech is also required. One of these contexts includes language-level factors[[63]](#footnote-63). Phonological weakness[[64]](#footnote-64) [[65]](#footnote-65) and mispronunciation phenomena[[66]](#footnote-66) [[67]](#footnote-67)suggest that a higher level of speech processing affects the perception of phonemes. In other words, in automatic speech recognition, the pronunciation lexicon and the language model jointly constrain the acoustic model to eliminate confusions within the acoustic model as much as possible, with the pronunciation lexicon eliminating acoustic confusions at the lexical level and the language model eliminating them at the shallow grammatical level.

Fourth, since the pronunciation lexicon provides candidate words for speech recognition, then on the other hand, the words in the pronunciation lexicon determine the choice of modeling units for the language model. As mentioned above, the process of modeling N-metric grammatical language models suffers from data sparsity, i.e., the distribution of N-metric grammars in the corpus is unbalanced, leading to inaccurate estimation of model parameters. Since the pronunciation lexicon limits the units of the language model, then the data sparsity problem of language model modeling can be prevented by selecting appropriate lexical entries.

From the engineering point of view, the pronunciation dictionary has two main roles in automatic speech recognition systems:

(1) it limits the number of trisyllables. When modeling trisyllabic phonemes, there are not three times as many trisyllabic phonemes as there are N phonemes. Because of the phonological features or the limitations of the pronunciation lexicon, some triphonemes are not possible to use in reality. For example, in Chinese, there is no need to model three consonant phonemes. The number of triphonemes modeled is determined by the pronunciation lexicon, and the decision process is divided into two types: intra-word triphonemes and inter-word triphonemes. The latter is the commonly used method nowadays. All inter-word triphonemes can be obtained by simply combining the words in the pronunciation dictionary two by two.

(2) Limiting the internal search space of words during decoding. It can be imagined that if all the trisyllables relate to all the other trisyllables in the decoding of large vocabulary continuous speech recognition, the network to be searched will be extremely large. And the role of pronunciation dictionary is to limit the path of decoding to search within words, which can only constitute the path of word pronunciation.

From the above analysis, the pronunciation dictionary has an important position in large vocabulary continuous speech recognition, and its content and scale have a great influence on the performance of speech recognition. At the same time, the linguistic-based rule descriptions for various confusing speech phenomena can also be self-error screening and correction at the lexicon level to improve the accuracy of speech recognition.

In terms of the size of the pronunciation lexicon, on the one hand, the more words there are the more difficult speech recognition is, mainly because: first, the more words there are the more similar words, and the confusion between words increases, due to the discrimination. Second, since the words in the pronunciation dictionary are the units for modeling the language model, the more words there are, the more serious the data sparsity problem of the language model modeling is, resulting in the more inaccurate estimation of the language model parameters; third, with the increase of the number of words, the search operation makes the computational overhead grow rapidly. On the other hand, the performance of speech recognition decreases the smaller the number of words. The fewer the words, the weaker the constraint on the acoustic model to eliminate acoustic confusion, and as the number of words decreases, the more Out-Of-Vocabulary (OOV) words in speech recognition, which greatly affects the performance of recognition. And from the content of the pronunciation dictionary, how to determine the entries in the pronunciation dictionary needs to consider the pronunciation confusion between words and the data sparsity problem of language model modeling.

Decoder

The decoder (Decoder) is one of the cores of a speech recognition system. It incorporates acoustic mode model, language model and pronunciation dictionary to construct a search space and search a word sequence that optimally matches the input observed feature vector sequence as the recognition result[[68]](#footnote-68).

The search network is the carrier of the search space, and the decoder can be divided into dynamic extended decoder and static extended decoder according to the different ways of search network expansion during the decoding process. A typical dynamic extension decoder is the dynamic extension decoder based on a shared prefix tree[[69]](#footnote-69) [[70]](#footnote-70). Time-synchronized Viterbi-Beam search is the most common search algorithm for automatic speech recognition. However, some knowledge sources cannot be integrated into the time-synchronous search algorithm, and other time-asynchronous search algorithms and multiple search strategies are usually used, i.e., a simple model is first used to quickly obtain some possible candidates, such as N-best lists or Word Lattice[[71]](#footnote-71) [[72]](#footnote-72); then, a more refined model or other knowledge sources are introduced to select the final recognition results[[73]](#footnote-73) [[74]](#footnote-74) [[75]](#footnote-75).

1. General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China, General Technical Specification for Chinese Speech Recognition System of the State Standardization Administration of China [Z]. [↑](#footnote-ref-1)
2. Yang Xingjun, Chi Huisheng, etc. Digital processing of speech signals [M] Electronic Industry Press, 1995. [↑](#footnote-ref-2)
3. K.H. Davis, R. Biddulph, S. Balashek. Automatic recognition of spoken digitsJJ. The Journal of the Acoustical Society of America 1952.24 (6): 637-642 [↑](#footnote-ref-3)
4. H F. Olson, H Belar. Phonetic typewriter[J]. The Journal of the AcousticalSociety of America. 1956.28(6): 1071-1081 [↑](#footnote-ref-4)
5. J. W. Forgie,C.D. Forgie.Results obtained from a vowel recognition computer program[J]. The Journal of the Acoustical Society of America. 1959.31(11):1480-1489 [↑](#footnote-ref-5)
6. P Denes. The design and operation of the mechanical speech recognizer at university college London[J]. Journal of the British Institution of Radio Engineers. 1959,4(4):219-229 [↑](#footnote-ref-6)
7. D. B. Fry. Theoretical aspects of mechanical speech recognition[J]. Journal of the British Institution of Radio Engineers.1959,19(4):211-218 [↑](#footnote-ref-7)
8. J. W. Forgie,C.D. Forgie. Results obtained from a vowel recognition computer program[J]. The Journal of the Acoustical Society of America. 1959.31(11):1480-1489 [↑](#footnote-ref-8)
9. T K.Vintsyuk. Speech discrimination by dynamic programming[J]. Cybernetics and Systems Analysis. 1968, 4(1): 81-88 [↑](#footnote-ref-9)
10. D.R. Reddy. Computer recognition of connected speech[J]. The Journal of the Acoustical Society of America. 1967, 42(2): 329-347 [↑](#footnote-ref-10)
11. V M. Velichko, N.G. Zagoruyko. Automatic recognition of 200 words[J]. International Journal of Man-Machine Studies. 19702(3): 223-234 [↑](#footnote-ref-11)
12. D.H. Klatt. Review of the ARPA speech understanding project[J]. The Journal of the Acoustical Society of America. 1977.62(6): 1345-1366 [↑](#footnote-ref-12)
13. G.M.White. Speech recognition: a tutorial overview[J]. Computer. 19769(5): 40-53 [↑](#footnote-ref-13)
14. F. Jelinek, L. Bahl, R. Mercer. Design of a linguistic statistical decoder for the recognition of continuous speech[J]. IEEE Transaction on Information Theory. 1975,21(3):250-256 [↑](#footnote-ref-14)
15. D.H. Klatt. Review of the ARPA speech understanding project[J]. The Journal of the Acoustical Society of America.1977.62(6):1345-1366 [↑](#footnote-ref-15)
16. L.D. Erman. Overview of the HEARSAY speech understanding research[J].ACM SIGART Bulletin.1976.6(56):9-16 [↑](#footnote-ref-16)
17. B. Lowerre. The HARPY speech understanding system[J]. Readings in SpeechRecognition.1990:576-586 [↑](#footnote-ref-17)
18. HSakoe. Two-level DP-matching -A dynamic programming-based pattern matching algorithm for connected word recognition[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing. 1979,27(6):588-595 [↑](#footnote-ref-18)
19. J.S. Bridle, M D Brown, R.M Chamberlain. An algorithm for connected word recognition[C]. Automatic Speech Analysis and Recognition. 1982,191-204 [↑](#footnote-ref-19)
20. C.Myers, L. Rabiner. A level building dynamic time warping algorithm for connected word recognition[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing.1981,29(2):284-297 [↑](#footnote-ref-20)
21. CH Lee, L.Rabiner. A frame-synchronous network search algorithm for connected word recognition[J]. IEEE Transactions on Acoustics, Speech, and SignalProcessing.1989,37(11):1649-1658 [↑](#footnote-ref-21)
22. S. K. Das, MA. Picheny. Issues in practical large vocabulary isolated word recognition: the IBM tangora system[M]. Automatic Speech and Speaker Recognition, Springer US,1996 [↑](#footnote-ref-22)
23. L.Rabiner,S.E.Levinson, A. E. Rosenberg,J.G. Wilpon. Speaker independent recognition of isolated words using clustering techniques[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing.1979.27(4):336-349 [↑](#footnote-ref-23)
24. R.P Lippmann. An introduction to computing with neural nets[J]. IEEE ASSPMagazine.1987,4(2):4-22 [↑](#footnote-ref-24)
25. A.Waibel,T. Hanazawa, G.Hinton, K.Shikano, K.J Lang. Phoneme recognition using time-delay neural networks[J] . IEEE Transactions on Acoustics Speech, and Signal Processing.1989,37(3):328-339 [↑](#footnote-ref-25)
26. A.J. Robinson. L. Almeida. J-M Boite, H Bourlard, et al. A neural network based, speaker independent, large vocabulary, continuous speech recognition system: The WERNICKE project[C]. Proceedings of the European Conference on Speech Communication and Technology.1993,1941-1944 [↑](#footnote-ref-26)
27. G. Zavaliagkos, Y. Zhao, R. Schwartz, J. Makhoul. A hybrid segmental neural net/hidden Markov model system for continuous speech recognition[J]. IEEE Transaction on Acoustics, Speech, and Signal Processing.1994.2:151-160 [↑](#footnote-ref-27)
28. K. F. Lee, H.W.Hon, R. Reddy. An overview of the SPHINX speech recognition system[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing1990,38(1):35-45 [↑](#footnote-ref-28)
29. Y. L. Chow,M.O. Dunham, O. A. Kimball, M. A. Krasner, G. F Kubala, JMakhoul, P.J.Price, S.Roucos,R.M.Schwartz. BYBLOS: The BBN continuous speech recognition system[C]. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.1987,89-92 [↑](#footnote-ref-29)
30. R Schwartz.Y Chow 0. Kimball. S.Roucos, M Krasner J.Makhoul. Context dependent modeling for acoustic phonetic recognition of continuous speech(C7Proceedings of the IEEE International Conference on Acoustics. Speech, and Signal Processing.1985,1205-1208 [↑](#footnote-ref-30)
31. B.H. Juang, S. Levinson, M. Sondhi. Maximum likelihood estimation for multivariate mixture observations of Markov chains[J]. IEEE Transaction on Information Theory.1986,32(2):307-309 [↑](#footnote-ref-31)
32. L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition[J].Proceedings of the IEEE.989.77(2):257-286 [↑](#footnote-ref-32)
33. P. F. Brownand, P. V. deSouza, R.L. Mercer, V. J. D. Petra, J. C. Lai. Class based n-gram models of natural language[J]. Computational Linguistics. 199218(4):467-479 [↑](#footnote-ref-33)
34. R.Kneser, H. Ney. Improved clustering techniques for class-based statistical language modeling[C]. Proceedings of the European Conference on Speech Communication and Technology.1993,21-23 [↑](#footnote-ref-34)
35. R. Kneser. Statistical language modeling using a variable context length(C]Proceedings of International Conference on Spoken Language.1996.494-497 [↑](#footnote-ref-35)
36. T. R.Niesler, P. C.Woodland. A variable-length category-based n-gram language model[C]. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.1996,164-167 [↑](#footnote-ref-36)
37. M-H. Siu, M. Ostendorf Variable n-gram and extensions for conversational speech language modeling[J]. IEEE Transactions on Speech and Audio Processing.2000,8(1):63-75 [↑](#footnote-ref-37)
38. R. Rosenfeld. A maximum entropy approach to adaptive statistical language modeling[J]. Computer Speech and Language. 1996,10(3):187 - 228 [↑](#footnote-ref-38)
39. R.Lau, R.Rosenfeld, S.Roukos. Trigger-based language models: A maximum entropy approach[C]. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.1993.45-48 [↑](#footnote-ref-39)
40. S. Khudanpur,J. Wu. A maximum entropy language model integrating n-grams and topic dependencies for conversational speech recognition/Cl. Proceedings of the IEEE International Conference on Acoustics. Speech, and Signal Processing, 1999 [↑](#footnote-ref-40)
41. R. Kuhn. R. D. Mori. A cache-based natural language model for speech reproduction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence1990.12(6):570-583 [↑](#footnote-ref-41)
42. S.J. Young,G. Evermann,D. Kershaw,G. Moore,et al. The HTK bookhttp://htk.eng.cam.ac.uk/.[Z] [↑](#footnote-ref-42)
43. M. Mohri, F. Pereira, M. Riley. Weighted finite-state transducers in speech recognition[J]. Computer Speech and Language.2002,16(1):69-88 [↑](#footnote-ref-43)
44. G.E. Dahl,D.Yu, L. Deng, A. Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition[J]. EEE Transactions on Acoustics, Speech, and Signal Processing.2012,20(1):30-42 [↑](#footnote-ref-44)
45. Y. Zhang, A. S. Hildebrand, S. Vogel. Distributed language modeling for n-best list re-ranking[C]. Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2006 [↑](#footnote-ref-45)
46. A. Emami,K.Papineni, J. Sorensen. Large-scale distributed language modeling[C]. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.2007 [↑](#footnote-ref-46)
47. Z. Chen,K-F. Lee,J.M Jing. Discriminative training on language model Proceedings of the International Conference on Spoken Language Processing2000 [↑](#footnote-ref-47)
48. H. Kuo, E. Forsler-Lussier, H. Jiang, C.-H. Lee. Discriminative training of language models for speech recogintion[C]. Proceeding of the IEEE International Conference on Acoustic, Speech, Signal Processing.2002 [↑](#footnote-ref-48)
49. B. Roark,M.Saraclar, M Collins. Discriminative n-gram language modeling[J]. Computer Speech and Language.2007,2:373-392 [↑](#footnote-ref-49)
50. Z. Zhou, H. Meng. Rescasting and discriminative n-gram model as a pseudo conventional n-gram model for LVCSR[C]. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.2008 [↑](#footnote-ref-50)
51. Y Bengio. R.Ducharme. P Vincent, C.Jauvin, et al. A neural probabilistic language model[J].Journal of Machine Learning Research. 2003, 3:1137-1155 [↑](#footnote-ref-51)
52. R.Sarikaya,A. Emami, M. Affy, B.Ramabhadran. Continuous space language modeling techniques[C]. Proceedings of the EEE International Conference on Acoustics, Speech, and Signal Processing.2010 [↑](#footnote-ref-52)
53. T. Mikolov, M. Karafit, L.Burget,J.Cernocky, S. Khudanpur. Recurrent neural network based language model[C]. Interspeech.2010 [↑](#footnote-ref-53)
54. T Mikolov S. Kombrink, L. Burget, J. Cernocky, S. Khudanpur. Extensions of recurrent neural network language models(C]. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.2011 [↑](#footnote-ref-54)
55. S. B. Davis.P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences[J]. IEEF Transactions on Acoustics, Speech, and Signal Processing.1980.28:357-366 [↑](#footnote-ref-55)
56. H Hermansky. Perceptual linear predictive (PLP) analysis of speech[J]. Journal of the Acoustical Society of America. 1990,87(4):1738-1752 [↑](#footnote-ref-56)
57. A. Mohamed,G.E. Dahl, G. Hinton. Acoustic modeling using deep belief networks[J]. IEEE Transactions on Audio, Speech, and Language Processing.201220(1):14-22 [↑](#footnote-ref-57)
58. P Schwarz. Phoneme recognition based on long temporal context[D], PhD thesis, Faculty of Information Technology, Brno University of Technology, 2008 [↑](#footnote-ref-58)
59. L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Proceedings of the IEEE.989.77(2):257-286 [↑](#footnote-ref-59)
60. L.E.Baum,T.Petrie,G. Soules,N.Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains[J]. The Annals of Mathematical Statistics.1970,41(1):164-171 [↑](#footnote-ref-60)
61. LE. Baum. An equality and associated maximization technique in statistic estimation for probabilistic functions of Markov processes[J]. Inequalities. 1972, 3:1-8 [↑](#footnote-ref-61)
62. A.P Dempster,N.M. Laird, D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society. 1977, 39(1):1-38 [↑](#footnote-ref-62)
63. D.W.Carroll. Psychology of language[M]. Thomson Higher Education,2008 [↑](#footnote-ref-63)
64. R.M. Warren. Perceptual restoration of missing speech sounds[J]. Science 1970,167:392-393 [↑](#footnote-ref-64)
65. RM. Warren, R.P Warren. Auditory illusions and confusions[J]. Scientific American.1970:30-36 [↑](#footnote-ref-65)
66. R. A. Cole. Listening for mispronunciations: a measure of what we hear during speech[J]. Perception and Psychophysics.1973,13:153-156 [↑](#footnote-ref-66)
67. WD.Marslen-Wilson, A. Welsh. Processing interactions and lexical access during word recognition in continuous speech[J]. Cognitive Psychology.1978, 10:29-63 [↑](#footnote-ref-67)
68. X L. Aubert. An overview of decoding techniques for large vocabulary continuous speech recognition[J]. Computer Speech and Language. 2002, 16(1):89114 [↑](#footnote-ref-68)
69. A. J. Hunt, A. W. Black. Unit selection in a concatenative speech synthesis system using large database[C]. Proceeding of the IEEE International Conference on Acoustic,Speech,and Signal Processing.1996,373-376 [↑](#footnote-ref-69)
70. H.Ney, R.Haeb-Umbach, B.-H. Tran, M.Oerder. Improvements in beam search for 10000-word continuous speech recognition[C]. Proceeding of the IEEE International Conference on Acoustic, Speech, Signal Processing.1992.9-12 [↑](#footnote-ref-70)
71. S.Ortmanns.H Ney. X. Aubert. A word graph algorithm for large vocabulary continuous speech recognition[J]. Computer Speech and Language. 199711(1):43-72 [↑](#footnote-ref-71)
72. A. Sixtus.Across-word phoneme models for large vocabulary continuous speech recognition. PhD. thesis. Universitatsbibliothek.2003 [↑](#footnote-ref-72)
73. J.L.Gauvain, L. Lamel, M. Adda-Decker. Developments in continuous speech dictation using the ARPA WSJ task[C]. Proceeding of the IEEE International Conference on Acoustic, Speech, Signal Processing.1995,65-68 [↑](#footnote-ref-73)
74. H. Murveit,J. Butzberger, V Digalakis, M. Weintraub. Large-vocabulary dictation using SRI's DECIPHER speech recognition system: Progressive search techniques(C]. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.1993,319-322 [↑](#footnote-ref-74)
75. P C.Woodland, J. J. Odell, V Valtchev, S.J. Young. Large vocabulary continuous speech recognition using HTKIC]. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.1994.125-128 [↑](#footnote-ref-75)