УДК 004.896
MSC: 68T10; 68T50; 68T42
DOI: 10.35330/1991-6639-2020-6-98-20-33

MODERN PROBLEMS OF AUTOMATIC SPEECH RECOGNITION

I.A. GURTUEVA

Institute of Computer Science and Problems of Regional Management –
Branch of Federal public budgetary scientific establishment «Federal scientific center
«Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences»
360000, KBR, Nalchik, 37-a, I. Armand St.
E-mail: iipru@rambler.ru

This paper provides a concise review of the most applied methods in speech recognition. Various principles of transcription developed in the Linguistic Data Consortium are discussed. The problems in evaluating the human level of efficiency in solving the problem of speech recognition are described. The typical errors made by a human are analyzed. It has been shown that transcribers demonstrate a high level of consistency with accurate transcription of pre-prepared English speech and fast transcription of conversational telephone speech. It is also shown that with increasing complexity of speech, the word disagreement rate increases. The results of a comparative analysis of errors generated by the speech system and those made by humans are presented. Their similarities and differences are analyzed. The modern automatic speech recognition problems are listed, the prospects for their solution and the directions of future research are estimated.

Keywords: deep learning, artificial intelligence, artificial neuron networks, speech recognition, human parity.

REFERENCES

1. Campbell M., Hoane A.J., Hsu F.-h. Deep Blue // Artificial intelligence. 2002. Vol. 134. Pp. 57-83.

2. Silver D., Huang A., Maddison C. J.,Guez A., Sifre L., Van Den Driessche G., Schrittwieser J., Antonoglou I., Panneershelvam V., Lanctot M., et al. Mastering the game of Go with deep neural networks and tree search // Nature. 2016. Vol. 529. Pp. 484-489.

3. Amodei D., Anubhai R., Battenberg E., Case C., Casper J., Catanzaro B., Hen J., Chrzanowski M., Coates A., Diamos G., et al. Deep Speech 2: End-to-end speech recognition in English and Mandarin // arXiv preprint arXiv:1512.02595. 2015.

4. Kristjansson T.T., Hershey J.R., Olsen P.A., Rennie S.J., Gopinath R.A. Super-human multi-talker speech recognition: the IBM 2006 Speech Separation Challenge system // Proc. Interspeech. 2006. Vol. 12. P. 155.

5. Weng C., Yu D., Seltzer M. L., Droppo J. Single-channel mixed speech recognition using deep neural networks // Proc. IEEE ICASSP. 2014. Pp. 5632-5636.

6. Pallett D.S. A look at NIST's benchmark ASR tests: past, present, and future // IEEE Automatic Speech Recognition and Understanding Workshop. 2003. Pp. 483-488.

7. Price P., Fisher W.M., Bernstein J., Pallett D.S. The DARPA 1000-word resource management database for continuous speech recognition // Proc. IEEE ICASSP. 1988. pp. 651–654.

8. Paul D.B., Baker J.M. The design for the wall street journal-based csr corpus // Proceedings of the workshop on Speech and Natural Language. 1992. Pp. 357-362.

9. *Graff D., Wu Z., MacIntyre R., Liberman M.* The 1996 broadcast news speech and language-model corpus // Proceedings of the DARPA Workshop on Spoken Language technology. 1997. Pp. 11-14.

10. Ljolje A. The AT&T 2001 LVCSR system // NIST LVCSR Workshop. 2001.

11. Philppov D. *Interaktivnoye golosovoye redaktirovaniye teksta s pomoshch'yu novykh rechevykh tekhnologiy ot Yandeksa* [Interactive Voice Text Editing Using New Speech Technologies from Yandex]. https://habr.com/ru/company/yandex/blog/243813/. 2014.

12. Chen S.F., Kingsbury B., Mangu L., Povey D., Saon G., Soltau H., Zweig G. Advances in speech transcription at IBM under the DARPA EARS program // IEEE Trans. Audio, Speech, and Language Processing. 2006. Vol. 14. Pp. 1596-1608.

13. Seide F., Li G., Yu D. Conversational speech transcription using context-dependent deep neural networks // Proc. Interspeech. 2011. Pp. 437-440.

14. Matsoukas S., Gauvain J.-L., Adda G., Colthurst T., Kao C.-L., Kimball O., Lamel L., Lefevre F., Ma J.Z., Makhoul J., et al. Advances in transcription of broadcast news and conversational telephone speech within the combined ears bbn/limsi system // IEEE Transactions on Audio, Speech, and Language Processing. 2006. Vol. 14. Pp. 1541-1556.

15. Stolcke A., Chen B., Franco H., Gadde V. R. R., Graciarena M., Hwang M.-Y., Kirchhoff K., Mandal A., Morgan N., Lei X., et al. Recent innovations in speech-to-text transcription at SRI-ICSI-UW // IEEE Transactions on Audio, Speech, and Language Processing. 2006. Vol. 14. Pp. 1729-1744.

16. Gauvain J.-L., Lamel L., Schwenk H., Adda G., Chen L., Lefevre F. Conversational telephone speech recognition // Proc. IEEE ICASSP. 2003. Vol. 1. Pp. 1-212.

17. Evermann G., Chan H. Y., Gales M. J. F., Hain T., Liu X., Mrva D., Wang L., Woodland P.C. Development of the 2003 CU-HTK conversational telephone speech transcription system // Proc. IEEE ICASSP. 2004. Vol. 1. Pp. 1-249. 2004.

18. Fry D.B. Theoretical aspects of mechanical speech recognition // J. British Inst. Radio Engr. 1959. Pp. 211-229.

19. Vintsyuk T.K. *Raspoznavaniye slov ustnoy rechi metodami dinamicheskogo programmi-rovaniya* [Speech discrimination by dynamic programming] // Kibernetika. 1968. 4 (2). Pp. 81- 88.

20. Viterbi A.J. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm // IEEE Trans. Information Theory, IT- 13. 1967. Pp. 260-269.

21. Reddy D.R. An approach to computer speech recognition by direct analysis of the speech wave // Tech. Report No. C549, Computer Science Dept., Stanford Univ. 1966.

22. Velichko V.M., Zagoruyko N.G. *Avtomaticheskoye raspoznavaniye ogranichennogo nabora ustnykh komand* [Automatic recognition of 200 words] // Int. J. Man- Machine Studies. 1970. 2. Pp. 223.

23. Sakoe H., Chiba S. Dynamic programming algorithm optimization for spoken word recognition // IEEE Trans. Acoustics, Speech, Signal Proc, ASSP-26 (1). 1978. Pp. 43-49.

24. Rabiner L. R., et. al. Speaker independent recognition of isolated words using clustering techniques // IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-27. 1979. Pp. 336-349.

25. Klatt D. Review of the ARPA speech understanding project // J.A.S.A. 1977. 62(6). Pp. 1324-1366.

26. Lowerre B. The HARPY speech understanding system // Trends in Speech Recognition, W. Lea, Ed., Speech Science Pub. 1990. Pp. 576-586.

27. Rabiner L. R., Juang B. H. Fundamentals of Speech Recognition. Prentice-Hall, Englewood Cliff. New Jersey. 1993.

28. Katagiri S. Speech pattern recognition using neural networks // W. Chou and B.-H. Juang (Eds.) Pattern Recognition in Speech and Language Processing, CRC Press. 2003. Pp. 115-147.

29. Myers C.S., Rabiner L.R. A level building dynamic time warping algorithm for connected word recognition // IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-29. 1981. Pp. 284-297.

30. Lee C.H., Rabiner L.R. A frame synchronous network search algorithm for connected word recognition // IEEE Trans. Acoustics, Speech, Signal Proc. 1989. 37 (11). Pp. 1649-1658.

31. Bridle J.S., Brown M.D. Connected word recognition using whole word templates // Proc. Inst. Acoust. Autumn Conf. 1979. Pp. 25-28.

32. Juang B.-H., Furui S. Automatic speech recognition and understanding: A first step toward natural human-machine communication // Proc. IEEE, 88, 8. 2000. Pp. 1142-1165.

33. Chou W. Mimimum classification error (MCE) approach in pattern recognition // Chou W., Juang B.-H. (Eds.) Pattern Recognition in Speech and Language Processing. CRC Press. 2003. Pp. 1-49.

34. Leggetter C.J., Woodland P.C. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models // Computer Speech and Language. 1995. 9. Pp. 171-185.

35. Varga A.P., Moore R.K. Hidden Markov model decomposition of speech and noise // Proc. ICASSP. 1990. Pp. 845-848.

36. Gales M. J. F., Young S.J. Parallel model combination for speech recognition in noise // Technical Report, CUED/FINFENG/ TR135. 1993.

37. Shinoda K., Lee C.H. A structural Bayes approach to speaker adaptation // IEEE Trans. Speech and Audio Proc. 2001. 9, 3. Pp. 276-287.

38. Stolcke A., Droppo. J. Comparing Human and Machine Errors in Conversational Speech Transcription. /Interspeech. 2017. Pp. 137-141. DOI: 10.21437.

39. Saon G., Kurata G., Sercu T., Audhkhasi K.,Thomas S., Dimitriadis D., Cui X., Ramabhadran B., Picheny M., Lim L.-L., Roomi B., Hall P. English Conversational Telephone Speech Recognition by Humans and Machines // INTERSPEECH 2017 DOI: 10.21437.

40. Lippmann R.P. Speech recognition by machines and humans // Speech Communication. 1997. Vol. 22. Issue 1. Pages 1-15. https://doi.org/10.1016/S0167-6393(97)00021-6.

41. Glenn M. L., Strassel S. M.,H. Lee, Maeda K., Zakhary R., Li X. Transcription Methods for Consistency, Volume and Efficiency // Proceedings of the International Conference on Language Resources and Evaluation, LREC. 2010. Malta.

42. Hannun A. Speech Recognition Is Not Solved. https://awni.github.io/speech-recognition. 2017.

Information about the author:

Gurtueva Irina Aslanbekovna, researcher of the Department of Computer Linguistics of the Institute of Computer Science and Problems of Regional Management of KBSC of the Russian Academy of Sciences.

360000, KBR, Nalchik, I. Armand street, 37-a.

E-mail: gurtueva-i@yandex.ru