Pengju YAN

Location: Shanghai, China
Phone: (86) 139-1803-6420
Email: yanpengju@gmail.com

Area of Interest	**Applied Machine Learning, Natural Language Processing, Large Language Models, Text-to-Speech, Speech Recognition, Data Mining or others related.**
Education	● Sep. 1998 – Jun. 2002, Department of Computer Science and Technology, **Tsinghua University**. **PhD.** ● Sep. 1993 – Jul. 1998, Department of Computer Science and Technology, **Tsinghua University**. Bachelor of Engineering.
Domain Expertise	● Proficiency in machine learning (ML), deep learning (DL), artificial intelligence (AI) and data mining algorithms. ● Extensive experiences in the following areas: natural language processing (NLP), speech synthesis, human-computer dialogue systems, large language model (LLM) applications, and AIDD (drug discovery).
Computer Skills	● Proficiency in Python/C/C++/C#/Java/Shell ● Proficiency in developing in Linux
Languages	● Mandarin: Mother tongue ● English: Fluent in reading, writing and speaking
Hangzhou Institute of Medicine, Chinese Academy of Sciences	● Aug. 2022 – , **Chief Machine Learning Expert** at the AIM Center. AIDD algorithm development: Aptamer optimization based on diffusion models (3-fold affinity improvement), RNA secondary structure prediction based on diffusion models, small molecule structure prediction based on equivariant neural networks and flow matching models (SOTA accuracy of < 0.20Å), etc. LLM application platform and research: Knowledge graph platform in biomedical field (36 million articles, 1.8 billion entities, 120 million relationships), intelligent search and paper reading platform based on generative agents (200 million scientific literature abstracts, 8 million full text), research on LLM+KG mixed model-based Q&A systems, etc.
Cyclone Robotics	● Mar. 2021 – Jun. 2022, **Algorithm VP**; team size 25. **Head of the Algorithm team.** Responsible for the application of AI technology in RPA, including CV and deep learning algorithms for application understanding, document understanding, and industrial application scenarios. Developed and delivered a CV-based UI element capture system, an OCR-based document understanding system, and a CV surveillance system for oil work platform. Planned and built the next-generation semantics-based product framework for RPA.
Tongdun	● Mar. 2018 – Jul. 2020, **Principal Data Scientist**; team size 25. **Algorithm Head of Intelligent Interaction**. Built the team from scratch. The R&D area of the team includes speech recognition, speaker recognition, text-to-speech, natural language understanding, and text mining. Leading the development of the conversational systems which is a company-level strategic product. **Conversational systems**: The average daily incoming volume of the conversational bot achieved 100,000 and gained a revenue of 10 million. Accuracy of ASR, sound quality of TTS, and accuracy of dialog NLU are on par with that of iFlytek products. A lower hanging-up rate and a higher repayment commitment rate were achieved in the overdue collection scenario.
Microsoft (Suzhou)	● Sep., 2016 – Jul. 2017, **Principal Applied Scientist**; team size 9. Algorithm architect of intelligent auto reply system. Conducted research of auto-reply by sequence-to-sequence deep learning modeling, and email highlighting by CDSSM (Convolutional Deep Structured Semantic Models). Managed the speech output back-end modeling team, working on efficient voice font building pipeline, and voice quality improvement on show case voices. Conducted research on prosody modeling using dual deep learning frameworks, and singing synthesis with RNN.
Alipay	● May, 2014 – May, 2016, **Senior Expert (P8)**; team size 13 for a period. Alipay Wallet (Mobile App) recommendation systems architecture design, including public account recommendations and operation targeting. Guiding team members in their day-to-day prediction and recommendation work. Algorithm R&D for the intelligent customer services, including account recognition, problem recognition, and operator assignment. A couple of KPIs were improved by more than 10% by his team. Applied Deep Learning to transaction trust model (detecting questionable transactions) with moderate improvement on accuracy and helped using of large amount of raw features.
eBay	● May, 2013 – Apr., 2014, **MTS 2 Software Engineer**. Designed new features to boost ranking accuracy based on detected fraudulent selling data. Ranking performance was improved significantly. Machine learning-based bot/botnet detection: Classification & clustering algorithms on the IP, user agent, and session levels, respectively. 10% more bots events were detected by our new algorithms.
AdSame	● Sep., 2011 – Apr., 2013, **General Manager of RTB Systems, Director of Algorithms**; team size 5. R&D of the fundamental algorithms for online advertising, in the scope of Natural Language Processing, Data Mining and Machine Learning. Research on topic model-based webpage content analysis and user profile (interest) generation

	based on user tracking, achieved 2-8 times of CTR by applying the models in content/user match-based high accuracy delivery. User demographic (gender, age, income, etc) recognition based on classification and regression models, used in Demographic-based high accuracy delivery.
SNDA	● Jul., 2009 – Aug., 2011, **Senior Researcher**; team size 9. Developed first class TTS (Text-to-Speech) system, with the quality on par with iFlytek. For the front-end components of word segmentation, named entity recognition, POS tagging, and prosody predictions, we adopted CRF to achieve high prediction accuracy. We also took measures to effectively reduce the decoding time and storage cost. Training duration, f0 and spectrum (LSF) models using HTK/HTS toolkits and developed high performance parameter generation algorithms. Under storage and computation constraints in embedded systems, we compacted storage structures and converted floating-point arithmetic to fixed-point.
Microsoft (Beijing)	● Jun., 2006 – Jun., 2009, Text-to-Speech development team, **SDE II**. Conducted the development of prosody models and acoustic models in TTS (Text-to-Speech) systems. Multi-layered CART f0 model + duration model: 44% preference gain. Energy prediction and filtering: 20% preference gain. Established spec of language expansion of prosody annotation and guided the annotation in German and Japanese. Yes-no-question script selection and recording: 70% preference gain for yes-no-questions. Applied HTK/HTS toolkit in the acoustic model training, including duration, F0 and LSF models.
Panasonic	● Aug., 2002 – Jun., 2006, **Team Lead**; team size 6. Design of the search algorithm of a LVCSR system, and the development of an embedded spoken language translation system. Development of a DTV program recommendation system, which made TV program recommendation to users according to their viewing behavior. Development of interoperation between networked home appliances, based on HTTP/HTML/UPnP/XML. Development of a SVM-based Chinese named entity recognition system. A One-Class SVM-based rejection scheme was proposed to reject unreliable candidates and achieved a high accuracy.
Tsinghua University	● Feb., 2000 – Jun., 2002, During the PhD. candidate period, being the chief designer of the airline information dialogue system of EasyFlight, he constructed the system architecture and built the natural language understanding component of which. ● Aug., 1999 – Jan., 2000, He participated in designing the campus navigation dialogue system of EasyNav, and built the stochastic word-class model and knowledge library of EasyNav.
Publication	[[12] Y. Yang, X. Fang, Z. Cheng, **P. Yan**, X. Li , "EquiBoost: An Equivariant Boosting Approach to Molecular Conformation Generation." *Proceedings of the AAAI Conference on Artificial Intelligence*. 2025. (**correspondence**) (rebuttal) [11] Z. Wang, Y. Feng, Q. Tian, Z. Liu, **P. Yan**, X. Li, "RNADiffFold: Generative RNA Secondary Structure Prediction using Discrete Diffusion Models." *Briefings in Bioinformatics* 26.1 (2025): bbae618. (**correspondence**) [10] Z. Wang, Z. Liu, W. Zhang, Y. Li, Y. Feng, S. Lv, H. Diao, Z. Luo, **P. Yan**, M. He, and X. Li "AptaDiff: de novo design and optimization of aptamers based on diffusion models." *Briefings in Bioinformatics* 25.6 (2024): bbae517. (**correspondence**) [9] F. Yang, S. Yang, P. Zhu, **P. Yan** and L. Xie, "Improving mandarin end-to-end speech synthesis by self-attention and learnable gaussian bias", [C]//2019 IEEE automatic speech recognition and understanding workshop (ASRU). IEEE, 2019: 208-213. [8] 吴悦, **燕鹏举**, 翟鲁峰. 基于二元背景模型的新词发现[J]. 清华大学学报：自然科学版, 2011, 51(9):4. [7] **Yan, Pengju**, and Fang Zheng. "Context directed speech recognition in dialogue systems." International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages. 2004. [6] **Yan P.J.**, Zheng F., Sun H., and Xu M.X., "Spontaneous Speech Parsing in Travel Information Inquiring and Booking Systems", Journal of Computer Science and Technology, Vol. 17, No. 6, November 2002. [5] Zheng, T. F., **Yan, P.**, et al. "Collection of a chinese spontaneous telephone speech corpus and proposal of robust rules for robust natural language parsing." Joint International Conference of SNLP-OCOCOSDA, Hua Hin, Thailand. 2002. [4] **Yan P.J.**, Zheng F., and Xu M.X., "Robust Parsing in Spoken Dialogue Systems", 7th European Conference on Speech Communication and Technology, In *INTERSPEECH* (pp. 2149-2152) [3] **燕鹏举**, 陆正中, 邬晓钧,等. 航班信息系统 EasyFlight[C]// 第六届全国人机语音通讯学术会议论文集. 2001. [2] **燕鹏举**, 郑方. 口语对话系统中的词类概率模型和知识表示[J]. 清华大学学报：自然科学版, 2001, 41(1):4. [1] **燕鹏举**. 对话系统中的词类模型,知识库管理及鲁棒的语言理解初探[D]. 清华大学.