DOI:10.1145/3481625

Several companies are trying push automatic speech recognition and other technologies past their current limitations.

BY JIA JIA, WEI CHEN, KAI YU, XIAODONG HE, JUN DU, AND HEUNG-YEUNG SHUM

The Practice of Speech and Language Processing in China

ALTHOUGH GREAT PROGRESS has been made in automatic speech recognition (ASR), significant performance degradation still exists in very noisy environments. Over the past few years, Chinese startup AISpeech has been developing very deep convolutional neural networks (VDCNN),[21] a new architecture the company recently began applying to ASR use cases.

Different than traditional deep CNN models for computer vision, VDCNN features novel filter designs, pooling operations, input feature map selection, and padding strategies, all of which lead to more accurate and robust ASR performance. Moreover, VDCNN is further extended with adaptation, which can significantly alleviate the mismatch between training and testing. Factor-aware training and cluster-adaptive training are explored to fully utilize the environmental variety and quickly adapt model parameters. With this newly proposed approach, ASR systems can improve the system robustness and accuracy, even in under very noisy and complex conditions.[1]

JD AI Research (JD), based in Beijing, China, has also made progress in auditory perception,

Figure 1. Models for sound event detection and localization.

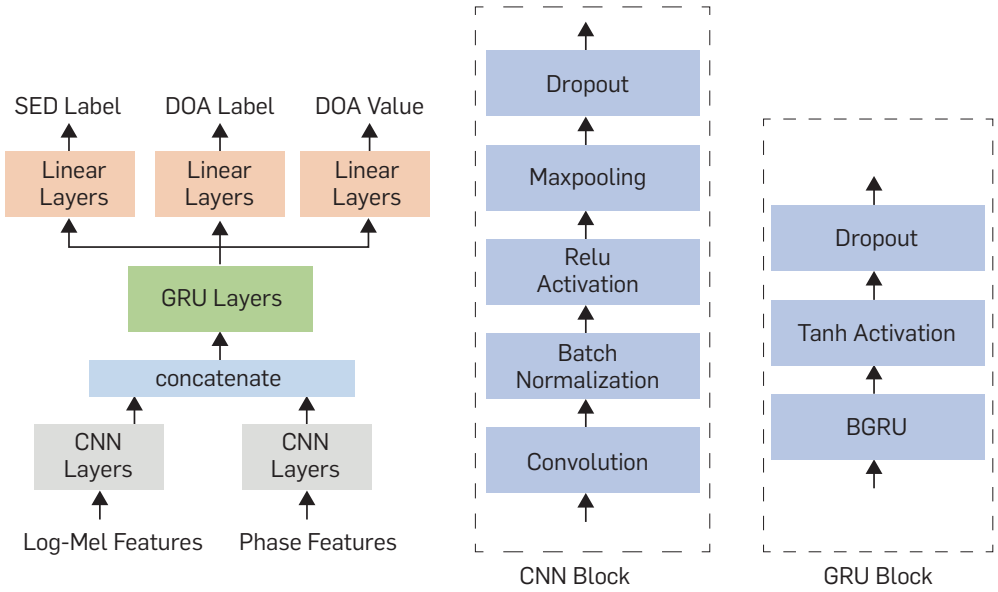


Figure 2. System diagram of densely connected multi-stage model for real-time speech enhancement.

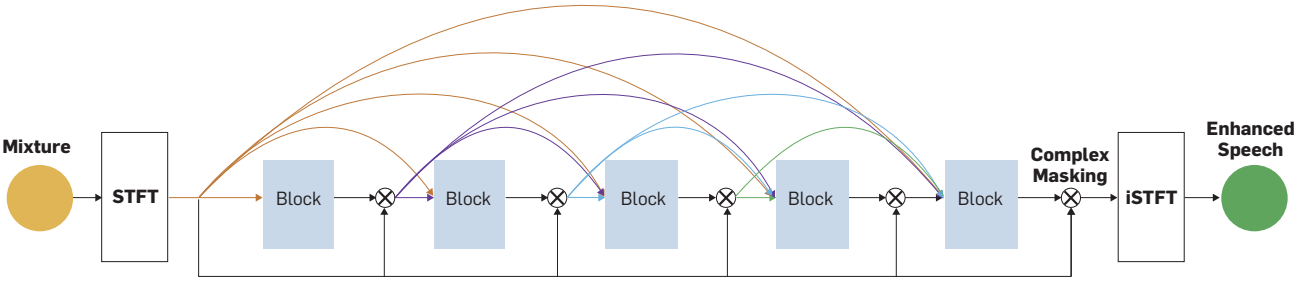
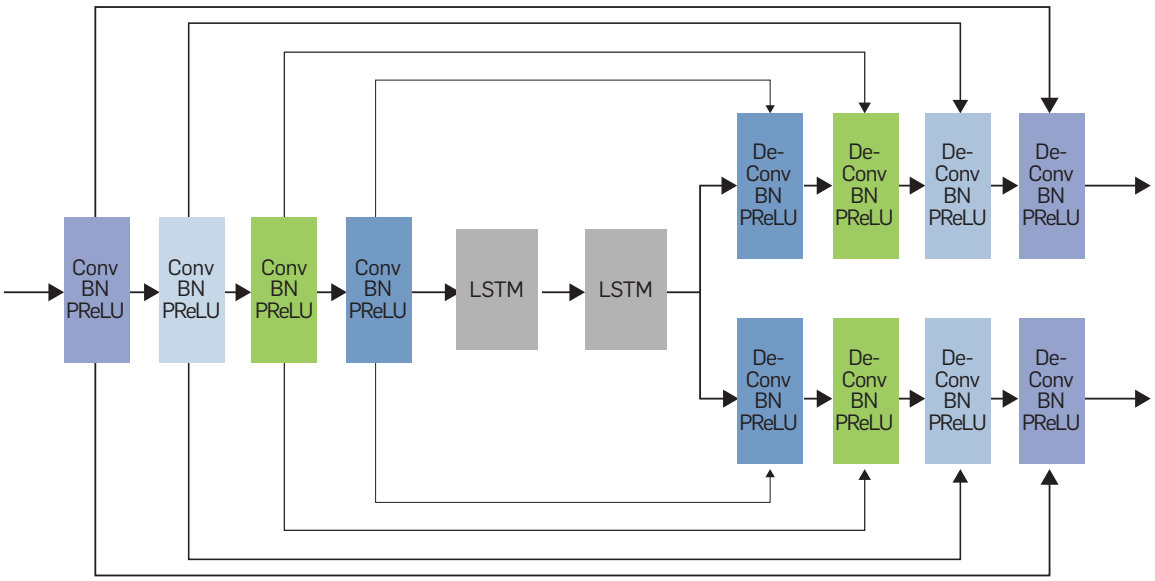


Figure 3. Block of the system.



aiming to detect and localize sound events, enhance target signals, and suppress reverberation. This is important not only because it enhances signals for speech recognition, but also because such information can be used for better decision-making in subsequent dialog systems.

For sound-event detection, as shown in Figure 1, a multi-beamforming-based approach is proposed: the diversified spatial information for the neural network is extracted using beamforming towards different directions.[32] For speech dereverberation, optimal smoothing-factor-based preprocessing is used to obtain a better presentation for the dereverberation network.[10] Beamforming and speech dereverberation are also used to generate augmented data for multichannel far-field speaker verification.[22] In terms of speech enhancement, neural Kalman filtering (KF) is proposed to combine conventional KF and speech evolution in an end-to-end framework.[31]

JD also ranked third in both the sound event localization and detection task of DCASE 2019 Challenge, and the FFSVC 2020 Challenge for far-field speaker verification.

For real-time speech enhancement, China-based Internet company Sogou proposes a deep complex convolution recurrent network (DCCRN) with restricted parameters and latency.[9] Different from real-valued networks, DCCRN adopts the complex CNN, complex long short-term memory (LSTM), and complex batch normalization layers, which are better suited for processing complex-valued spectrograms. Moreover, as shown in Figure 2 and Figure 3, a computational, efficient, real-time speech-enhancement network is proposed with densely connected, multistage structures.[11] The model applies sub-band decomposition and progressive strategy to achieve superior denoising performance with lower latency.

For end-to-end ASR, self-attention networks (SAN) in transformer-based architectures[23] show promising performance, so a transformer-based, attention-based encoder/decoder (AED) is selected as the base architecture.

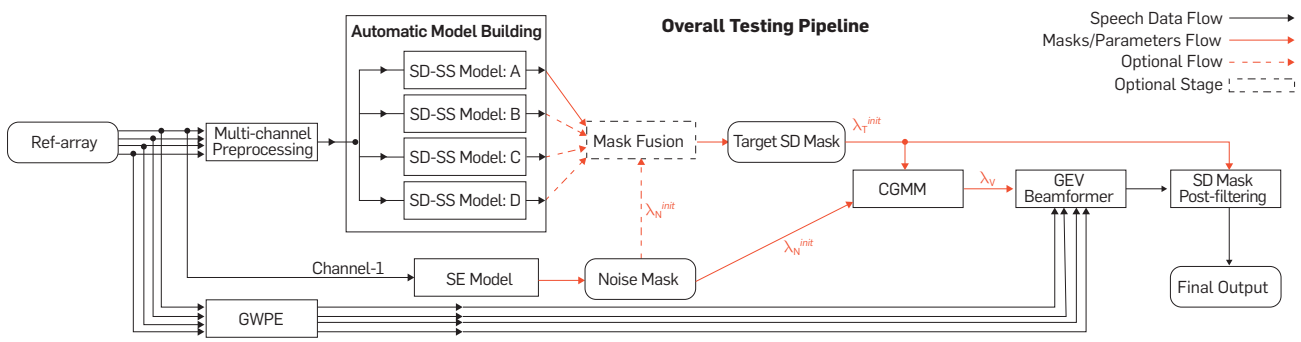
One approach is to improve AED performance for non-real-time speech transcription. Transformer-based architectures can easily achieve slightly better results than traditional hybrid systems in ordinary scenarios. However, transformer-based models collapse under some conditions, such as conversational speech and recognition of proper nouns. Relative positional embedding (RPE) and parallel scheduled sampling (PSS)[39] are adopted to improve generalization and stability. As transformer architecture is good at global modeling, and speech recognition relies more on local information, local modeling is further combined with CCNs and feedforward sequential memory networks (FSMN)[7] to the transformer to improve the modeling of local speech variance. To improve acoustic feature extraction of encoders, Sogou uses connectionist temporal classification (CTC) and cross entropy (CE), multitask joint training of the transformer. With this strategy, a 100,000-hour transformer achieves a 25% improvement compared to Kaldi-based hybrid systems.

A second research strategy is streaming AED. To that end, Sogou proposed an adaptive monotonic chunk-wise attention (AMoChA) mechanism,[6] which can adaptively learn chunk-length at each step to calculate context vectors for streaming attention. Transformer acoustic range is adaptively computed for each token in a streaming decoding fashion. For the CTC and CE joint-trained transformer, CTC output is viewed as first-pass decoding while the attention-based decoder is seen as second-pass decoding. Thus, the encoder is trained in a chunk-wise manner for streaming AED. This method is similar to non-auto-regressive decoding.[8]

The 100,000-hour streaming AED achieved a 15%–20% relative improvement compared to Kaldi-based hybrid streaming systems. Generally, ASR systems and speech enhancement (SE) systems are trained and deployed separately, because

Figure 4. The overall diagram of USTC-iFLYTEK front-end processing system for the CHiME-5 challenge.[20]

Here, SD-SS means speaker-dependent speech separation, SE model is a deep learning-based speech enhancement model, GWPE denotes the generalized weighted prediction error algorithm for dereverberation, CGMM means complex Gaussian mixture model, and GEV means generalized eigenvalue.



DCCRN adopts the complex CNN, complex long short-term memory (LSTM), and complex batch normalization layers, which are better suited for processing complex-valued spectrograms.

they typically have different purposes. Moreover, enhanced speech is detrimental to ASR performance. However, joint training of SE and ASR can significantly improve the performance of speech in high-noise environments while maintaining the performance of clean speech. For Sogou, the joint training system of the CRN-based SE model and the transformer-based ASR model results in an average relative improvement of 23% in noisy conditions and 5% in clean conditions.

Visual information is another way to boost speech recognition performance in noisy conditions. Google first proposed the Watch, Listen, Attend and Spell (WLAS) network, which jointly learns audio and visual information in the recognition task.[4] Sogou adopted a modality attention network based on WLAS[40] for adaptively integrating audio and visual

Figure 5. The embeddings for the future and past chunked sentences are concatenated to form the Cross Utterance (CU) context vector, which is concatenated with the phoneme encoder output vectors to form the input of the decoder.

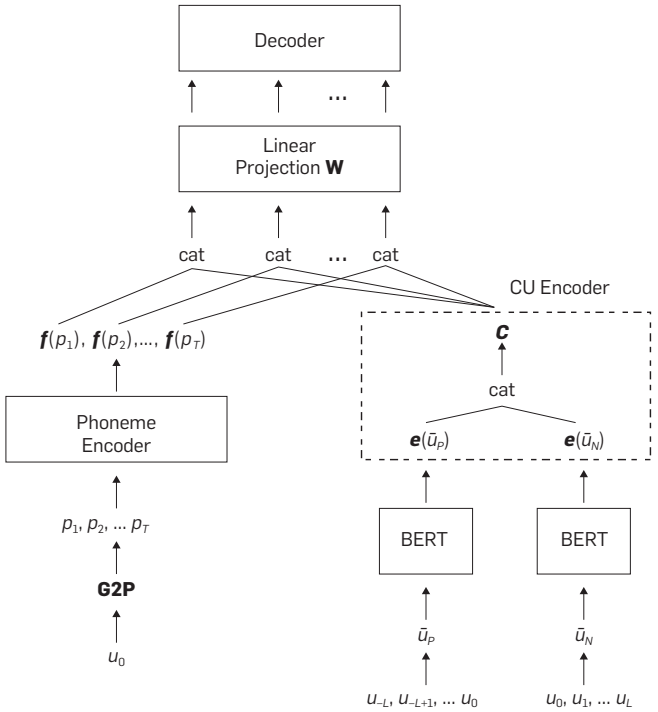
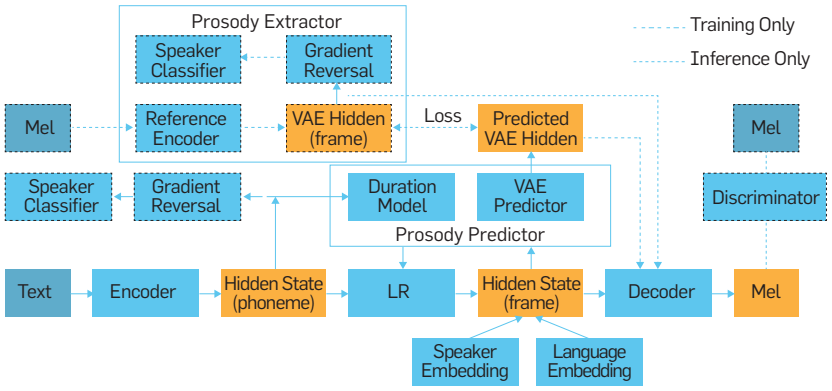


Figure 6. StyleTTS architecture.



information, which achieved a 35% performance improvement in 0-dB noisy conditions.

iFLYTEK, together with the National Engineering Laboratory for Speech and Language Information Processing at the University of Science and Technology of China (USTC), proposed novel, high-dimensional regression approaches to solve classical speech-signal preprocessing problems and is outperforming traditional methods by relaxing the constraints of many mathematical model assumptions.[5,20,29] The organization has finished in first place in several prestigious challenges, including all four tasks of the CHiME-5 speech recognition challenge,[20] two tasks of the CHiME-6 speech recognition challenge,[27] all tasks of the DIHARD-III Speech Diarization Challenge,[15] and the Sound Event Localization and Detection (SELD) task of the DCASE2020 Challenge.[13] These challenges, especially CHiME-5/6 and DIHARD-III, are quite relevant to common "cocktail party problems" found in real multi-speaker scenarios. Figure 4 shows an overview of the USTC-iFLYTEK front-end processing system for the CHiME-5 challenge.

Robust Speaker Identification

Deep learning-based methods have been widely applied in this research area, achieving a new milestone for speaker identification and anti-spoofing. However, it is still difficult to develop a robust speaker identification system under complex, real-world scenarios such as short utterance, noise corruption, and channel mismatch. To boost speaker verification performance, AISpeech proposes new approaches to achieve more discriminant speaker embeddings within two frameworks.

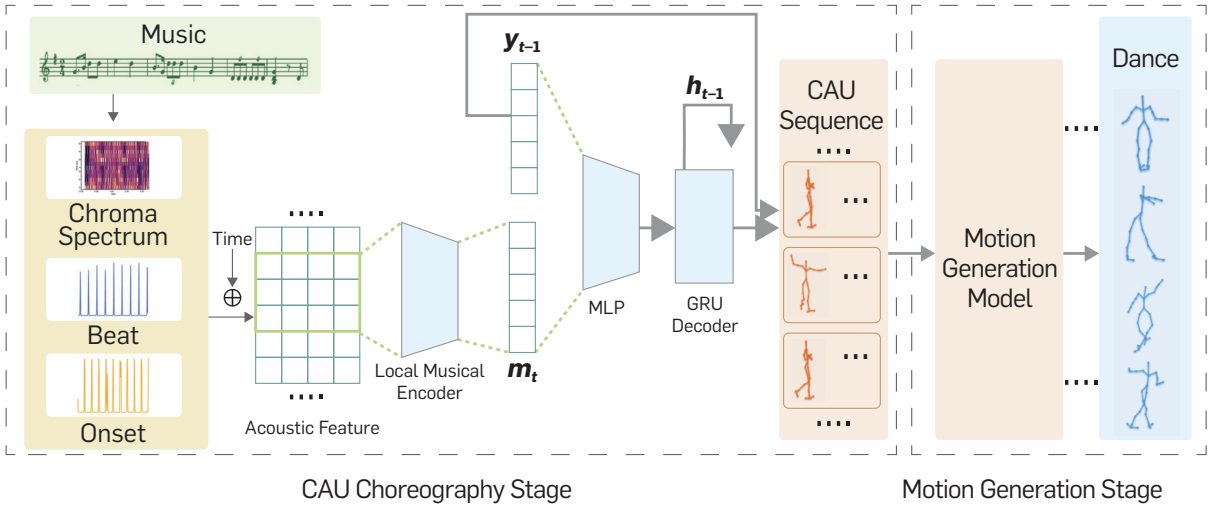
Within a cascade framework, a neural network-based deep discriminant analysis (DDA)[24,26] is suggested to project i-vector to more discriminant embeddings. The direct-embedding framework uses a deep model with more advanced center loss and A-softmax loss, and focal loss is also explored.[25] Moreover, traditional i-vector and neural embeddings are combined with neural network-based DDA to achieve another improvement. Furthermore, AISpeech proposes the use of deep generative models—for example, generative adversarial network (GAN) and variational autoencoder (VAE) models—to perform data augmentation directly on speaker embeddings, which would be used for robust probabilistic linear discriminant analysis (PLDA) training and to improve system accuracy.[2,34] With these newly proposed approaches, the speaker recognition system can significantly improve system robustness and accuracy under noisy and complex conditions.[3]

Robust TTS

To build robust and highly efficient TTS systems, research on both end-to-end network structures and neural vocoders was conducted. JD proposed an end-to-end speech synthesis framework—duration informed auto-regressive network (DIAN)[19]—which removes the attention mechanism with the help of a separate duration model. This eliminates common skipping and repeating issues. Efficient WaveGlow (EWG), a flow-based neural vocoder, was proposed in Song et al.[18] Compared with the baseline WaveGlow, EWG can reduce inference time cost by more than half, without any obvious reduction in speech quality. To study mixed lingual TTS systems, we look into speaker embedding and phoneme embedding, and study the choice of data for model training in Xue et al.[30] As shown in Figure 5, cross-utterance (CU) context vectors are used to improve the prosody generation for sentences in a paragraph in end-to-end fashion.[28]

Sogou also proposed an end-to-end TTS framework—Sogou-StyleTTS (see Figure 6)—to synthesize highly expressive voice.[12] For front-end text analysis, a cascaded, multitask BERT-LSTM model is adopted. And the acoustic model is improved over FastSpeech,[14] which is composed of a multilayer transformer encoder-decoder and a duration model. Hierarchical VAE is used to extract

Figure 7. The pipeline of the ChoreoNet.



VDCNN features novel filter designs, pooling operations, input feature map selection, and padding strategies, all of which lead to more accurate and robust ASR performance.

prosodic information unsupervised to decouple timbre and rhythm, which are considered as style, and a rhythm decoder, to predict the above prosody information. Using this structure, any timbre and rhythm can be combined to achieve style control and introduce GAN to further improve the sound quality, which brings the distribution of acoustic features closer to real voice. Finally, multiband MelGAN architecture[33] is proposed to invert the Mel spectrogram feature representation into waveform samples. Based on StyleTTS, a text-driven, digital-human generation system is proposed to realize a realistic digital human: a multi-modality, generative technology to model the digital human's voice, expressions, lips, and features jointly.

To generate more realistic facial expressions and lip movements, both face reconstruction and generative models are used to map from text to video frames. Moreover, to generate more expressive actions (Figure 7), Sogou cooperated with Tsinghua Tiangong Laboratory to carry out some exploratory work, such as creating digital-human music. ChoreoNet,[35] a two-stage music-to-dance synthesis framework, imitates human choreography procedures. The framework first devises a CAU prediction model to learn the mapping relationship between music and CAU sequences. Afterward, a spatial-temporal inpainting model is devised to convert the CAU sequence into continuous dance motions.

Network Compression

Faced with a need to deploy deep learning methods on edge devices, model compression without accuracy degradation has become a core challenge. Neural network language models (NNLM) have proven to be fundamental components for speech recognition and natural language processing in the deep learning era. Effective NNLM compression approaches that are independent of neural network structures are therefore of great interest. However, most compression approaches usually achieve a high compression ratio at the cost of significant performance loss. AISpeech proposes two advanced, structured-quantization techniques, namely product quantization[16] and soft binarization,[36] to enable the realization of a very high NNLM compression ratio compared to uncompressed models—70–100 without performance loss.[37] The diagram of product quantization for NNLM compression is shown in Figure 8.

Conclusion

These research outcomes have been widely used in many areas, including customer service, robotics, and smart home devices. For example, as shown in Figure 9, Xiaoice, originally developed at Microsoft in Beijing, now at XiaoBing.ai, is uniquely designed as an artificial intelligence companion with an emotional connection to satisfy the human need for communication, affection, and social belonging.[17,38] These techniques have successfully driven efficient, sustainable, and stable development, and aim to improve the future of the whole society. C

Figure 8. Diagram of product quantization for NNLM compression.

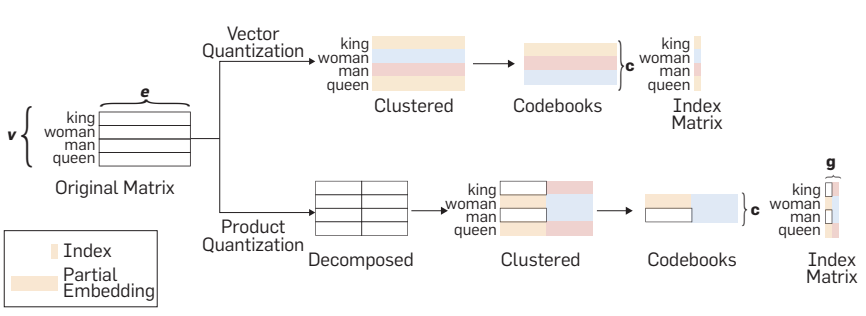
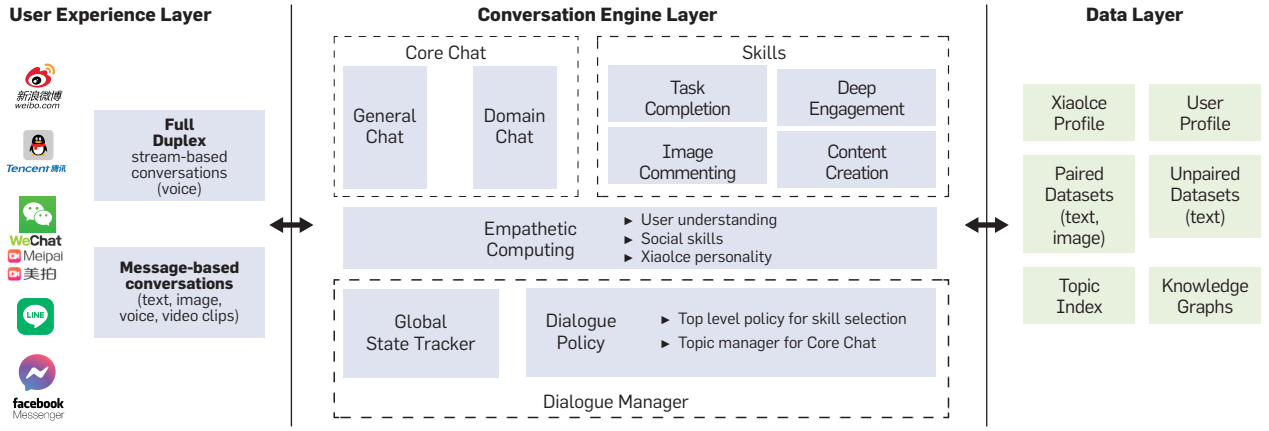


Figure 9. Xiaolce system architecture.



References

1. Bi, M., Qian, Y., and Yu, K. Very deep convolutional neural networks for LVCSR. In *Proceedings of the 16th Annual Conf. Intern. Speech Communication Assoc.*, 2015.
2. Chen, Z., Wang, S., and Qian, Y. Adversarial domain adaptation for speaker verification using partially shared network. In *Proceedings of Interspeech 2020*, 3017–3021.
3. Chen, Z., Wang, S., Qian, Y., and Yu, K. Channel invariant speaker embedding learning with joint multi-task and adversarial training. In *Proceedings of the IEEE 2020 Intern. Conf. Acoustics, Speech and Signal Processing*, 6574–6578.
4. Chung, J., Senior, A., Vinyals, O., and Zisserman, A. Lip reading sentences in the wild. In *Proceedings of the 2017 IEEE Conf. Computer Vision and Pattern Recognition*, 3444–3453.
5. Du, J., Tu, Y., Dai, L., and Lee, C. A regression approach to single-channel speech separation via high-resolution deep neural networks. *IEEE/ACM Trans. Audio, Speech, and Language Processing 24*, 8 (2016), 1424–1437.
6. Fan, R., Zhou, P., Chen, W., Jia, J., and Liu, G. An online attention-based model for speech recognition. In *Proceedings of Interspeech 2019*, 4390–4394.
7. Gao, Z., Zhang, S., Lei, M., and McLoughlin, I. SAN-M: Memory equipped self-attention for end-to-end speech recognition. In *Proceedings of Interspeech 2020*, 6–10.
8. Higuchi, Y., Watanabe, S., Chen, N., Ogawa, T., and Kobayashi, T. Mask CTC: Non-autoregressive end-to-end ASR with CTC and mask predict. In *Proceedings of Interspeech 2020*, 3655–3659.
9. Hu, Y. et al. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. (2020); arXiv:2008.00264.
10. Kothapally, V., Xia, W., Ghorbani, S., Hansen, J., Xue, W., and Huang, J. SkipConvNet: Skip convolutional neural network for speech dereverberation using optimally smoothed spectral mapping. (2020), arXiv:2007.09131.
11. Li, J. et al. Densely connected multi-stage model with channel wise sub-band feature for real-time speech enhancement. In *Proceedings of 2021 IEEE Intern. Conf. Acoustics, Speech and Signal Processing.*
12. Meng, F. et al. The Sogou system for Blizzard Challenge. In *Proceedings of 2020 Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge*, 49–53.
13. Politis, A., Adavanne, S., and Virtanen, T. Sound event localization and detection task. *2020 DCASE Challenge;* http://dcase.community/challenge2020/task-sound-event-localization-anddetection-results
14. Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T. FastSpeech: Fast, robust, and controllable text to speech. *NIPS* (2019), 3165–3174.
15. Ryant, N., Church, K., Cieri, C., Du, J., Ganapathy, S., and Liberman, M. The third DIHARD Speech Diarization Challenge; https://sat.nist.gov/dihard3#tab_leaderboard
16. Shi, K. and Yu, K. Structured word embedding for low memory neural network language model. In *Proceedings of Interspeech 2018*, 1254–1258.
17. Shum, H., He, X., and Li, D. From Eliza to XiaoIce: Challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering 19*, 1 (2018), 10–26.
18. Song, W., Xu, G., Zhang, Z., Zhang, C., He, X., and Zhou, B. Efficient WaveGlow: An improved WaveGlow vocoder with enhanced speed. In *Proceedings of Interspeech 2020*, 225–229.
19. Song, W., Yuan, X., Zhang, Z., Zhang, C., Wu, Y., He, X., and Zhou, B. Dian: Duration informed auto-regressive network for voice cloning. In *Proceedings of 2021 IEEE Intern. Conf. Acoustics, Speech and Signal Processing.*
20. Sun, L., Du, J., Gao, T., Fang, Y., Ma, F., and Lee, C. A speaker-dependent approach to separation of far-field multi-talker microphone array speech for front-end processing in the CHiME-5 Challenge. *IEEE J. Selected Topics in Signal Processing 13*, 4 (2019), 827–840.
21. Tan, T., Qian, Y., Hu, H., Zhou, Y., Ding, W., and Yu, K. Adaptive very deep convolutional residual network for noise robust speech recognition. *IEEE/ACM Trans. Audio, Speech, and Language Processing 26*, 8 (2018), 1393–1405.
22. Tong, Y. et al. The JD AI speaker verification system for the FFSVC 2020 Challenge. In *Proceedings of Interspeech 2020*, 3476–3480.
23. Vaswani, A. et al. Attention is all you need. (2017); arXiv:1706.03762.
24. Wang, S., Huang, Z., Qian, Y., and Yu, K. Discriminative neural embedding learning for short-duration text-independent speaker verification. *IEEE/ACM Trans. Audio, Speech, and Language Processing 27*, 11 (2019), 1686–1696.
25. Wang, S., Qian, Y., and Yu, K. Focal KL-divergence based dilated convolutional neural networks for co-channel speaker identification. In *Proceedings of 2018 IEEE Intern. Conf. Acoustics, Speech and Signal Processing*, 5339–5343.
26. Wang, S., Yang, Y., Wu, Z., Qian, Y., and Yu, K. Data augmentation using deep generative models for embedding based speaker recognition. *IEEE/ACM Trans. Audio, Speech, and Language Processing 28* (2020), 2598–2609.
27. Watanabe, S., Mandel, M., Barker, J., and Vincent, E. *The 6th CHiME Speech Separation and Recognition Challenge* (2020); https://chimechallenge.github.io/chime6/results.html
28. Xu, G., Song, W., Zhang, Z., Zhang, C., He, X., and Zhou, B. Improving prosody modelling with cross-utterance BERT embeddings for end-to-end speech synthesis. In *Proceedings of 2021 IEEE Intern. Conf. Acoustics, Speech and Signal Processing.*
29. Xu, Y., Du, J., Dai, L-R., and Lee, C-H. A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Trans. on Audio, Speech, and Language Processing 23*, 1 (2014), 7–19.
30. Xue, L., Song, W., Xu, G., Xie, L., and Wu, Z. Building a mixed-lingual neural TTS system with only monolingual data. In *Proceedings of Interspeech 2019* 2060–2064.
31. Xue, W., Quan, G. Zhang, C., Ding, G., He, X., and Zhou, B. Neural kalman filtering for speech enhancement. 2020; arXiv:2007.13962.
32. Xue, W., Tong, Y., Zhang, C., Ding, G., He, X., and Zhou, B. Sound event localization and detection based on multiple DOA beamforming and multi-task learning. In *Proceedings of Interspeech 2020*, 5091–5095.
33. Yang, G., Yang, S., Liu, K., Fang, P., Chen, W., and Xie, L. Multiband MelGAN: Faster waveform generation for high-quality text-to-speech. In *Proceedings of the 2021 IEEE Spoken Language Technology Workshop*, 492–498.
34. Yang, Y., Wang, S., Gong, X., Qian, Y., and Yu, K. Text adaptation for speaker verification with speaker-text factorized embeddings. In *Proceedings of the 2020 IEEE Intern. Conf. Acoustics, Speech and Signal Processing*, 6454–6458.
35. Ye, Z., Wu, H., Jia, J., Bu, Y., Chen, W., Meng, F., and Wang, Y. ChoreoNet: Towards music to dance synthesis with choreographic action unit. In *Proceedings of the 28th ACM Intern. Conf. Multimedia* (2020), 744–752.
36. Yu, K., Ma, R., Shi, K., and Liu, Q. Neural network language model compression with product quantization and soft binarization. *IEEE/ACM Trans. Audio, Speech, and Language Processing 28* (2020), 2438–2449.
37. Zhao, Z., Liu, Y., Chen, L., Liu, Q., Ma, R., and Yu, K. An investigation on different underlying quantization schemes for pre-trained language models. In *Proceedings of 2020 CCF International Conf. Natural Language Processing and Chinese Computing*. Springer, 359–371.
38. Zhou, L., Gao, J., Li, D., and Shum, H. The design and implementation of Xiaoice, an empathetic social chatbot. *Computational Linguistics 46*, 1 (2020), 53–93.
39. Zhou, P., Fan, R., Chen, W., and Jia, J. Improving generalization of transformer for speech recognition with parallel schedule sampling and relative positional embedding. (2019); arXiv:1911.00203.
40. Zhou, P., Yang, W., Chen, W., Wang, Y., and Jia, J. Modality attention for end-to-end audio-visual speech recognition. In *Proceedings of the 2019 IEEE Intern. Conf. on Acoustics, Speech and Signal Processing*, 6565–6569.

Jia Jia, Tsinghua University, Beijing, China.

Wei Chen, Sogou Corporation, Beijing, China.

Kai Yu, Shanghai Jiao Tong University, Shanghai, China.

Xiaodong He, JD AI Research, Beijing, China.

Jun Du, University of Science and Technology of China, Hefei, China.

Heung-Yeung Shum, XiaoBing.ai, Beijing, China.