陈 曦 | CHEN Xi

GitHub: https://github.com/OzymandiasChen

Phone: (+86) 19821250283

Email: chenxi19@fudan.edu.cn

EDUCATION

Hong Kong University of Science and Technology *Hong Kong, China*

Doctor of Philosophy, GPA: 3.7/4.3 *2023/09* – Present

Fudan University *Shanghai, China*

Master of Engineering in Computer Technology, GPA: 3.01/4 *2019/09 – 2022/01*

- Research Areas: Multimodal/Speech/Music Emotion Recognition, Singing Voice Detection

Dalian University of Technology *Dalian, Liaoning, China*

Bachelor of Engineering in Computer Science and Technology, GAP: 3.27/5, Ranking: 20/102; *2014/09 – 2018/07*

- Honors: National Encouragement Scholarship (2015), Learning Excellent Award (2015)

WORK EXPERIENCE

Baidu Inc. *Beijing, China*

Speech Algorithm Engineer *2022/02 – 2022/07*

- Customized Keyword Spotting algorithm development.
- Auto Keyword Spotting algorithm research and development.

Tencent Music Entertainment *Shenzhen, Guangdong, China*

Audio and Music Algorithm Intern *2021/06 – 2021/09*

- Singing Evaluation algorithm development (vocal range, pitch accuracy, vocal stability, and sense of rhythm).
- Accented Singing Identification algorithm research and development.

RESEARCH EXPERIENCE

Rule-Based Network for Multimodal Emotion Recognition *2022/11 – 2023/05*

Based on the relationship between audio and text, two rules of designing Multimodal Emotion Recognition model are proposed, and corresponding rule-based multimodal attention network (MMRBAN) is designed.

- As for the relationship between audio and text, since text modality is not independent of the audio modality, the former is embedded in the latter naturally, we propose Rule 1: The audio module should be more expressive than the text module. In MMRBAN, we adopt *M* Transformer encoder layers for the audio module, and *N* for the textual module, with restricting $M > N$.
- As for how the audio and the text contribute to emotion expression, since the emotion expressed in each modality can be homogeneous or heterogeneous to the overall emotion the speaker expressed, we propose Rule 2: Each single-modality emotion representation should be dynamically fused into the multimodal emotion representation. In MMRBAN, we design a Dynamic Weighting and Fusion Block, which is part of the modality interaction module.

Metric Learning with Time-domain Shifted-window Transformer for Accented Singing Identification

This work is the first dedicated in Accented Singing Identification, with three new challenges discussed and the corresponding Metric Learning with Time-domain Shifted-window Transformer (ML_TSWTF) proposed. *2021/06 – 2021/09*

- As for audio representation, the posterior probability graph of allophone is used, because it reveals the phoneme and the phonetic realization simultaneously and is robust to pitch and rhythm which are irrelevant in the singing accent.
- As for backbone, since audio representation is of high resolution and music pieces are usually in a length of a few minutes, a Transformer with time-domain shifted-window is designed, which is computationally efficient in both time and space.

- As for learning objective, since various unformal articulation methods contribute to various latent clustering centers in the accented singing, the triplet loss in metric learning, rather than the cross-entropy loss, is used, which doesn't minimize intra-class distance among negative samples.

Similarity-based Semi-supervised Learning Method for Singing Voice Detection *2020/09 – 2021/06*

Compared with previous methods, the Similarity-based Semi-supervised Learning Method for Singing Voice Detection (SSSL_SVD) can alleviate the data scarcity problem with no additional human labor, no requirement of prerequisite knowledge, and no missing critical polyphonic knowledge.

- To enrich the diversity of training data, the Self-training Semi-supervised Learning method is used, which has two benefits. One is that it is prerequisite-free, another is that it can mine information from real-world unlabeled data without missing critical polyphonic knowledge, such as the singing articulation, the synchronized nature between singing and accompaniment.
- To explore timber-related information and further facilitate decision-making, a similarity-based measurement, which is of higher entropy, is proposed.
- The proposed has a 2.2% promotion on accuracy and achieves comparable results with SOTA algorithms.

PUBLICATIONS

- **Xi Chen**. [Accepted] "MMRBN: Rule-Based Network for Multimodal Emotion Recognition". IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024.
- **Xi Chen**, Yongwei Gao, and Wei Li. "Singing Voice Detection via Similarity-based Semi-supervised Learning". ACM International Conference on Multimedia in Asia (ACM MM Asia), 2022.
- Shuai Yu, **Xi Chen**, and Wei Li. "Hierarchical Graph-based Neural Network for Singing Melody Extraction". IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022.
- **Xi Chen**, Lei Wang, and Wei Li." Channel-Wise Attention Mechanism in Convolutional Neural Networks for Music Emotion Recognition". Proceedings of the 8th Conference on Sound and Music Technology (CSMT), 2020.
- **Xi Chen**. [Position Paper, Under review] "Social-PTM: Pre-training Model for Social Media Language Understanding using Hierarchical Contextual Constructive Learning Framework". The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING), 2024.
- Shuai Yu, Yi Yu, **Xi Chen**, and Wei Li. HANME: hierarchical attention network for singing melody extraction. IEEE Signal Processing Letters, vol. 28, pp. 1006-1010, 2021.

SKILLS

Languages: English (TOEFL: 97), Chinese (Native).

Programming Languages: Python, Shell, C++, MATLAB.