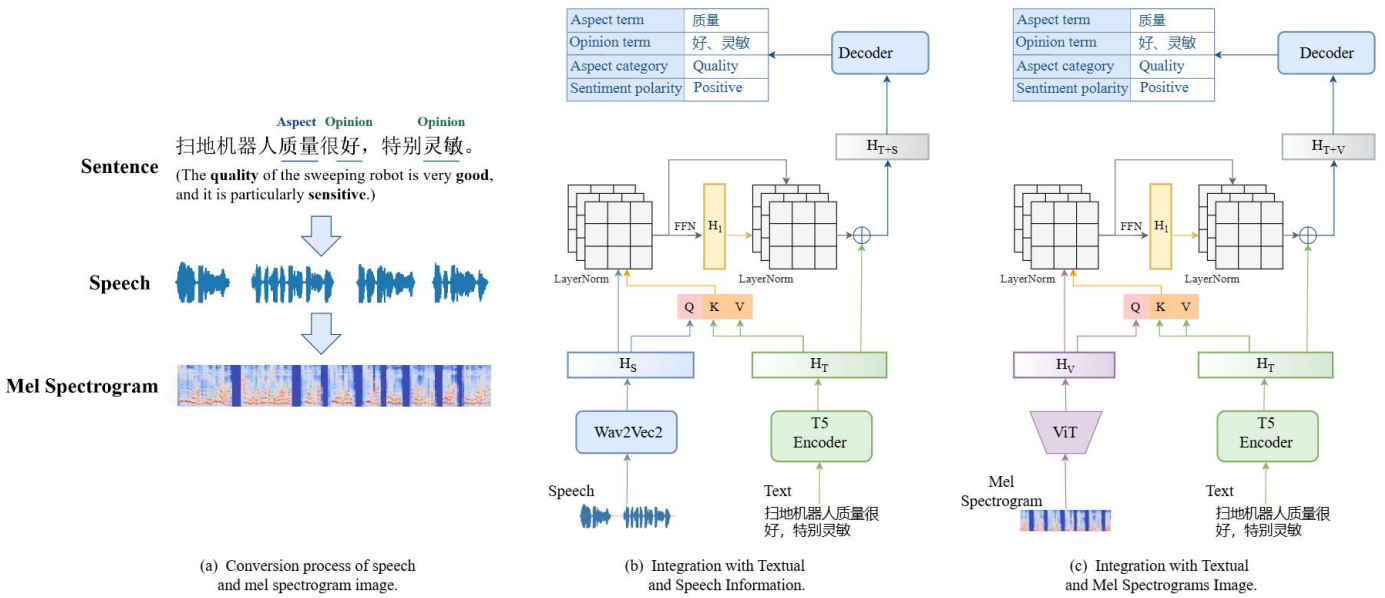
Exploring Speech Clues for Chinese Aspect-based Sentiment Analysis

Haowei LIU, Ye WANG, Xiaotong JIANG, Zhongqing WANG, Guodong ZHOU

Frontiers of Computer Science, DOI: 10.1007/s11704-025-50062-3

Problems & Ideas

- Problems of conventional Aspect-Based Sentiment Quadruple Extraction:
  - Traditional single-text-based methods underperform in sentiment quadruple extraction.
  - Multimodal Chinese quadruple extraction datasets are scarce, and audio features' distinct advantages remain unexplored.
- Ideas: We construct a large-scale text-audio dataset that integrates text and audio modalities to exploit their complementary roles, thereby enhancing the performance of aspect-sentiment-opinion-target quadruple extraction.



(a) Conversion process of speech and mel spectrogram image.

(b) Integration with Textual and Speech Information.

(c) Integration with Textual and Mel Spectrograms Image.

Speech and Mel Spectrogram Conversion Process and Model Framework.

Main Contributions

- Contributions:
  - We construct a large-scale human-annotated Chinese ABSA dataset with textual and acoustic contents that offers a valuable resource for advancing research in sentiment analysis and related fields.
  - We investigated the impact of two acoustic representations (raw audio signals and Mel-spectrograms) on model performance and developed a multimodal fusion framework to integrate text and acoustic modalities.
  - Experimental results show that acoustic-enhanced multimodal models outperform all baselines on our dataset, validating audio integration's effectiveness in sentiment analysis.

Table 1 Comparison with baselines. Bold numbers denote the best performance; underlined numbers indicate the second-best results.

Method	Sweeping Robot (R.)			Game Console (G.)			Accessories (A.)		
	P.	R.	F1.	P.	R.	F1.	P.	R.	F1.
EClassify	0.4324	0.5034	0.4652	0.2565	0.4143	0.3168	0.1623	0.2813	0.2058
GAS	0.7362	0.7293	0.7327	0.4593	0.4322	0.4453	0.3928	0.3366	0.3625
Seq2Path	0.7302	0.7356	0.7329	0.4325	0.4497	0.4409	0.3572	0.3675	0.3623
OTG	0.7358	0.7329	0.7343	0.4387	0.4516	0.4451	0.3602	0.3627	0.3614
UAUL	0.7425	0.7372	0.7398	0.4598	0.4351	0.4471	**0.3935**	0.3367	0.3629
LLaMA	0.7038	0.7195	0.7116	0.4167	0.4316	0.4240	0.3546	0.3442	0.3493
BLIP2	0.7455	0.7346	0.7400	0.4492	0.4383	0.4437	0.3692	0.3565	0.3627
InternLM-VL	0.7216	0.7279	0.7247	0.4360	**0.4823**	0.4580	0.3273	0.3224	0.3248
BERT-M3T	0.4727	0.5036	0.4877	0.3268	0.3442	0.3353	0.2305	0.2468	0.2384
Ours(T+S)	0.7604	0.7403	0.7502	**0.4712**	0.4521	0.4615	0.3649	**0.3791**	0.3719
Ours(T+I)	**0.7627**	**0.7430**	**0.7527**	0.4698	0.4659	**0.4679**	0.3901	0.3761	**0.3830**

Comparison with baselines. Bold numbers denote the best performance; underlined numbers indicate the second-best results. T+S means text-speech fusion and T+I means text-spectrogram fusion.