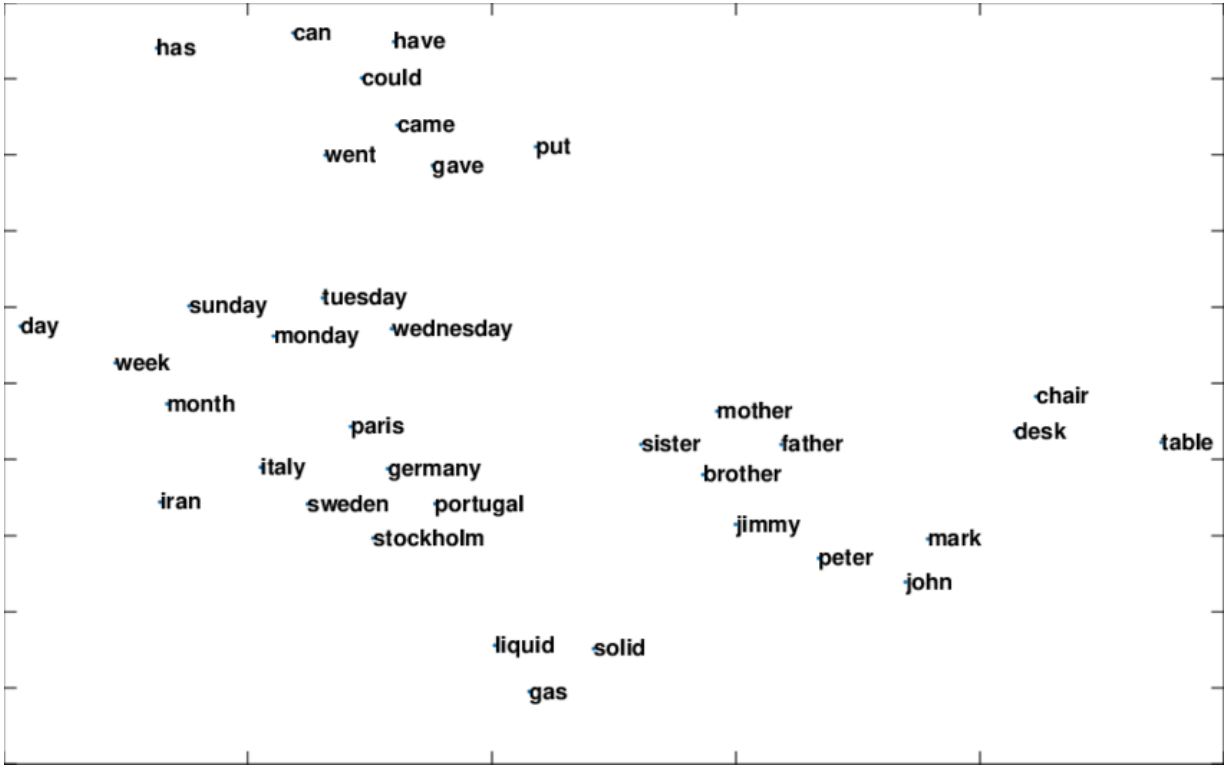
A Multimodal Music Emotion Classification Method Based On Multi feature Combined Network Classifier

Changfeng Chen, Qiang Li, Mathematical Problems in Engineering, 2020
Institute of Intelligent and Software Technology, Hangzhou Danzi University, China

경영과학연구실 이태헌
2023.02.01

Q&A

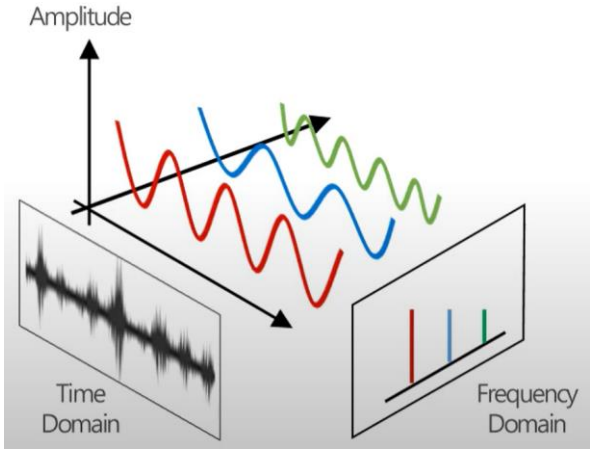
- Wordembedding 결과물 2차원?
- PCA 2차원 차원축소



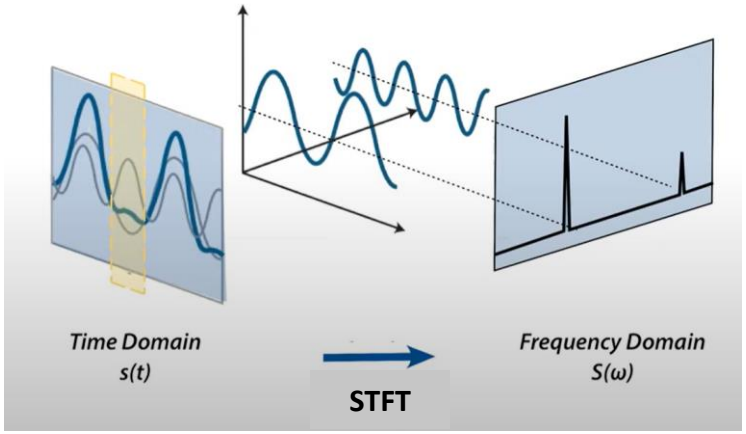
Audio data (STFT & Spectrogram)

- Time domain represents audio features over time
- Frequency domain expresses characteristics in terms of various frequencies that make up audio
- STFT(Short time fourier transform) takes a Fourier transform at a certain time of the audio and lists them in time order
- The spectrogram can be obtained through the STFT transformation

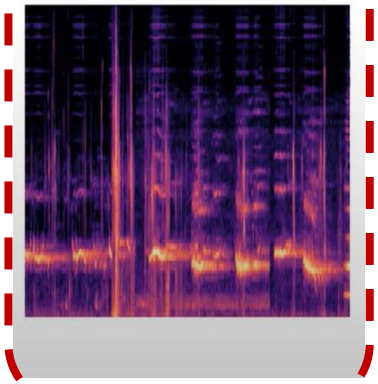
Time domain & Frequency Domain



STFT (Time to Frequency)



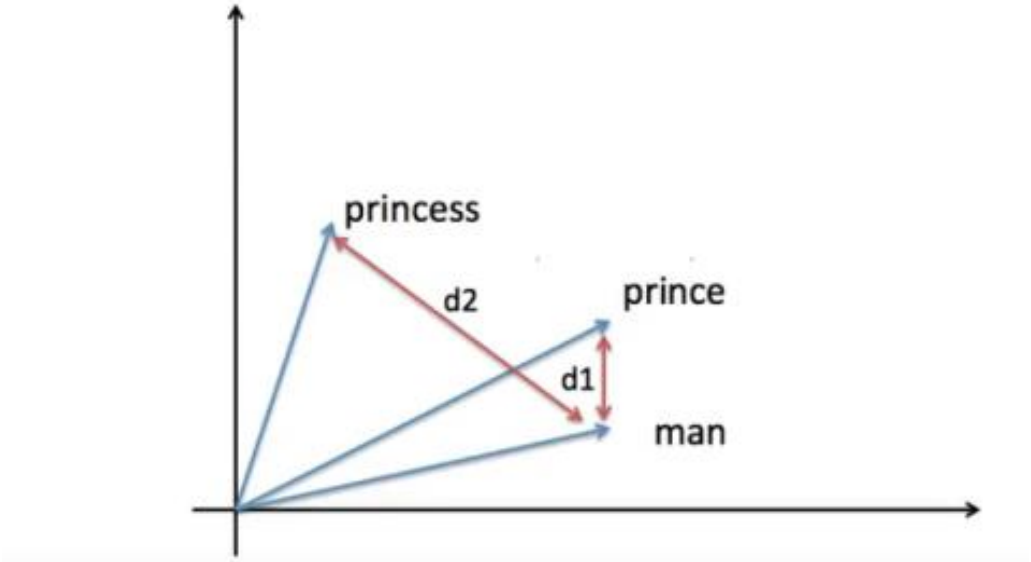
Spectrogram



Word2vec

- The lyrics are the words of the music
- Word2vec turns words into vectors
- In order to calculate the similarity between words, the similarity can be expressed by vectorizing a low-dimensional vector into a multi-dimensional space
- Word2vec assumes that words appearing in similar positions have similar meanings

Word2vec example

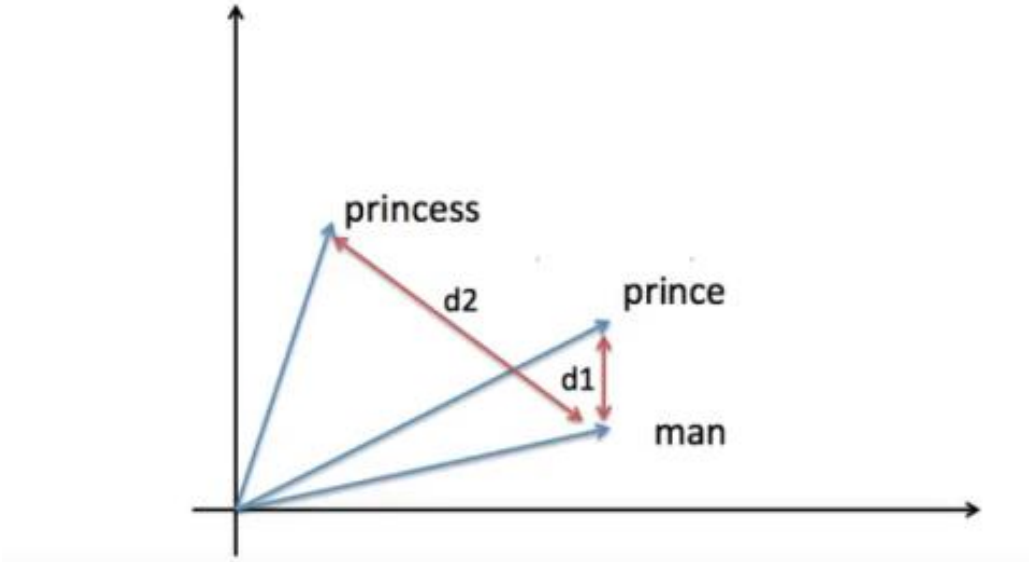


d1 is shorther than d2, man is similar to prince than princess

Word2vec

- The lyrics are the words of the music
- Word2vec turns words into vectors
- In order to calculate the similarity between words, the similarity can be expressed by vectorizing a low-dimensional vector into a multi-dimensional space
- Word2vec assumes that words appearing in similar positions have similar meanings

Word2vec example



d1 is shorther than d2, man is similar to prince than princess

The challenge of MER(Music emotion recognition)

- Emotions are subjective to each person and change depending on the meaning situation.
- Many researchs have been attempted to solve the problem by taking a multi-modal approach (Audio feature + a (Lyrics, Album image, Symbolic feature, biological data etc..))

Results of MER task in MIREX

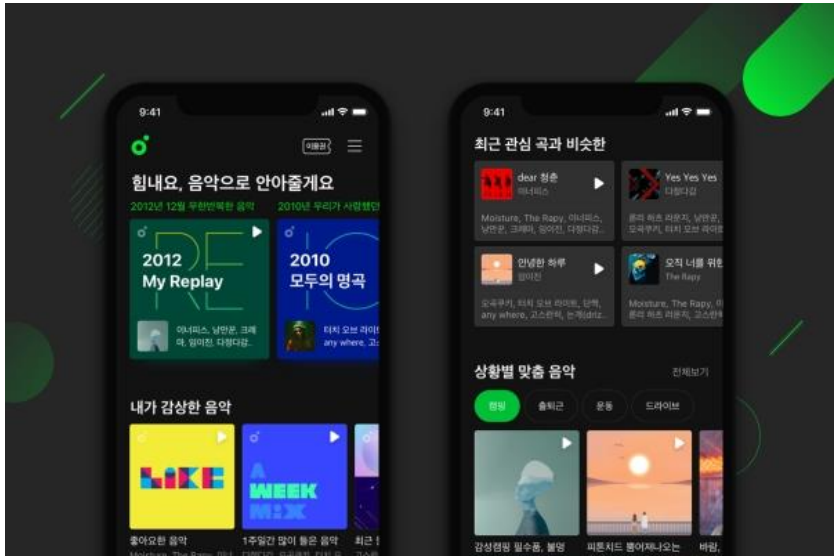
Year	Method	Accuracy/%
2020	Mel spectrogram + CNN	69.5
2019	-	68
2018	STFT + CNN	61.17
2017	Mel spectrogram + DCNN+SVM	69.83
2016	FFT, MFCC + CNN	63.33
2015	-	66.17
2014	MFCC + SVM	66.33
2013	Visual and acoustic features + SVM	68.33
2012	Audio features + SVM based models	67.83
2011	Audio features + SRC	69.5

MiREX (Music Information Retrieval Evaluation eXchange)

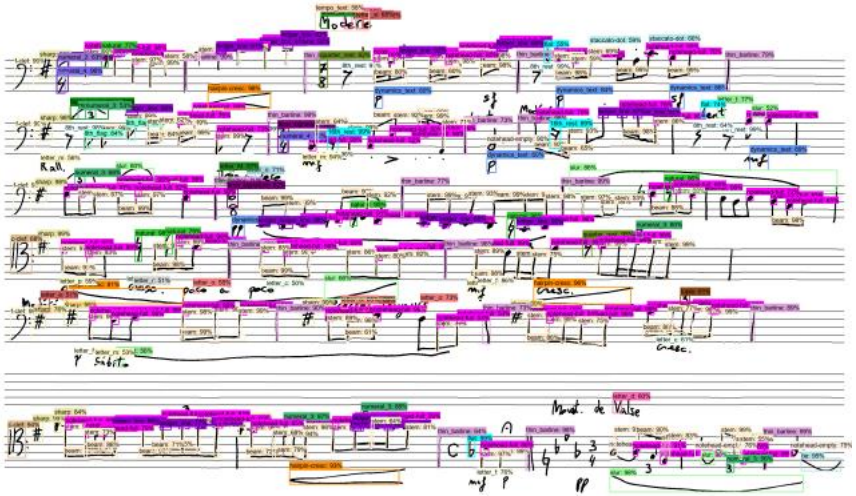
Application of MER(Music emotion recognition)

- MER can be widely used in many music fields
  (Music recommendation, Music retrieval, Music visualization, automatic music composing, Psychotherapy)

Music recommendation



Automatic music composing



Related works

Music Emotion classification research of audio

- *R. G. Ramani and K. Priya, "Improvised emotion and genre detection for songs through signal processing and genetic algorithm," Concurrency and Computation: Practice and Experience, 2019*
- *Y. C. Lin, Y. H. Yang, and H. H. Chen, "Exploiting online music tags for music emotion classification," ACM Transactions on Multimedia Computing, Communications, and Applications, 2011*

Music emotion classification research of Lyrics

- *X. Chen and T. Y. Tang, "Combining content and sentiment analysis on lyrics for a lightweight emotion-aware Chinese song recommendation system," in Proceedings of the 2018 10th International Conference on Machine Learning and Computing, pp. 85–89, Macau, China, February2018.*
- *C. S. Lee, M. H. Wang, L. C. Chen et al., "Fuzzy semantic agent based on ontology model for Chinese lyrics classification," in Proceedings of the 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 4254–4259, IEEE, Miyazaki, Japan, October 2018.*

Music emotion classification research of Multi-modal

- *F. H. Rachman, R. Sarno, and C. Fatichah, "Music emotion classification based on lyrics-audio using corpus based emotion," International Journal of Electrical and Computer Engineering (IJECE), vol. 8, no. 3, p. 1720, 2018.*
- *W. Shi and S. Feng, "Research on music emotion classification based on lyrics and audio," in Proceedings of the 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), pp. 1154–1159, IEEE, Chongqing, China, October 2018.*
- *E. Jokinen, R. Saeidi, T. Kinnunen, and P. Alku, "Vocal effort compensation for MFCC feature extraction in a shouted versus normal speaker recognition task," Computer Speech & Language, vol. 53, pp. 1–11, 2019.*

Problem statement & Key idea

Problem statement

This paper proposes a multimodal music emotion classification method. To solve the music classification problem with a multi-modal approach using audio & lyrics feature

Key idea

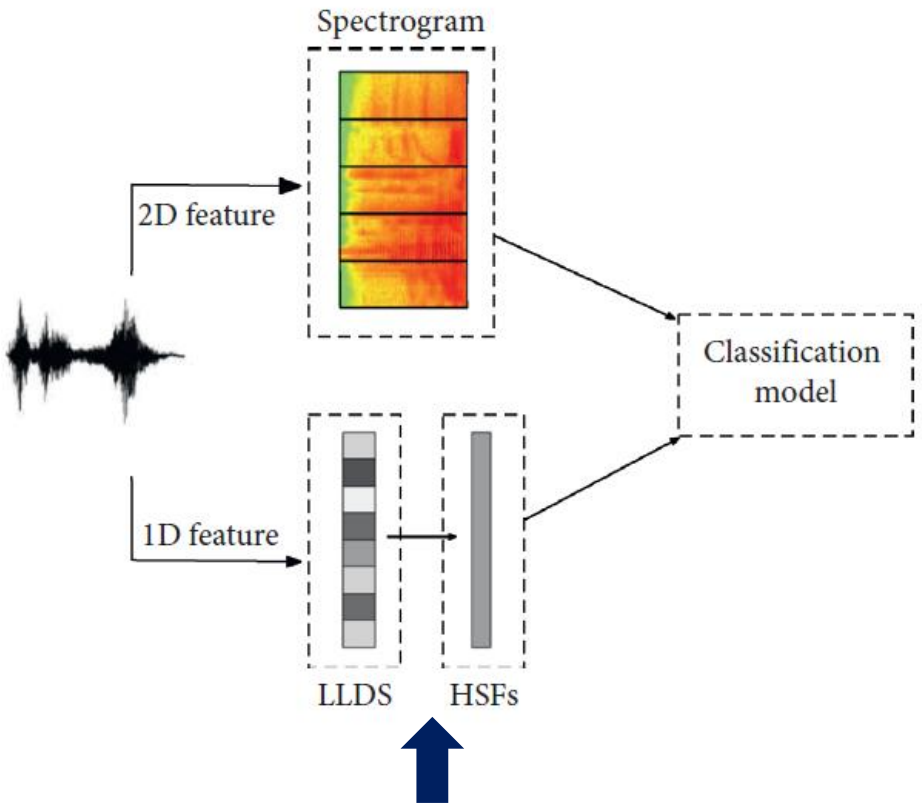
1. Pre-processing by separating human voice and background music from audio

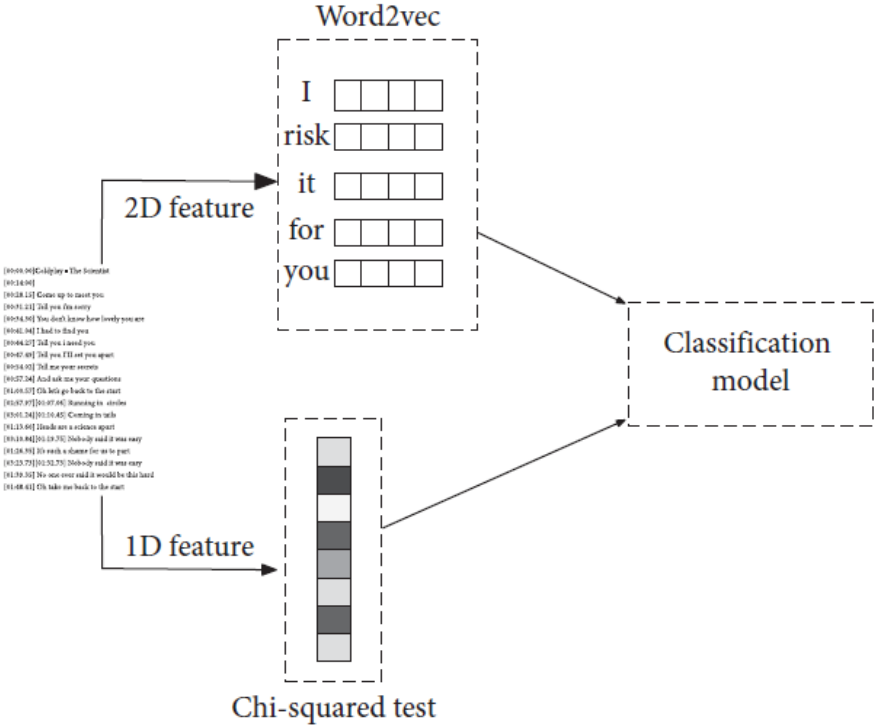
2. Proposed 2D + CNN-LSTM, 1D + DNN model architecture
   - Audio feature : 2D Spectrogram, 1D LLD (Low level descriptor, ex) MFCC, spectral feature
   - Lyrics feature : 2D Word embedding, 1D Word frequency vector

3. Audio & lyrics feature fusion with Stacking Ensemble structure

Audio, Lyrics classification input

- Audio Feature : Spectrogram(2D), LLD(1D) feature
- Lyrics Feature : Word embedding(2D), Word frequency vector(1D)

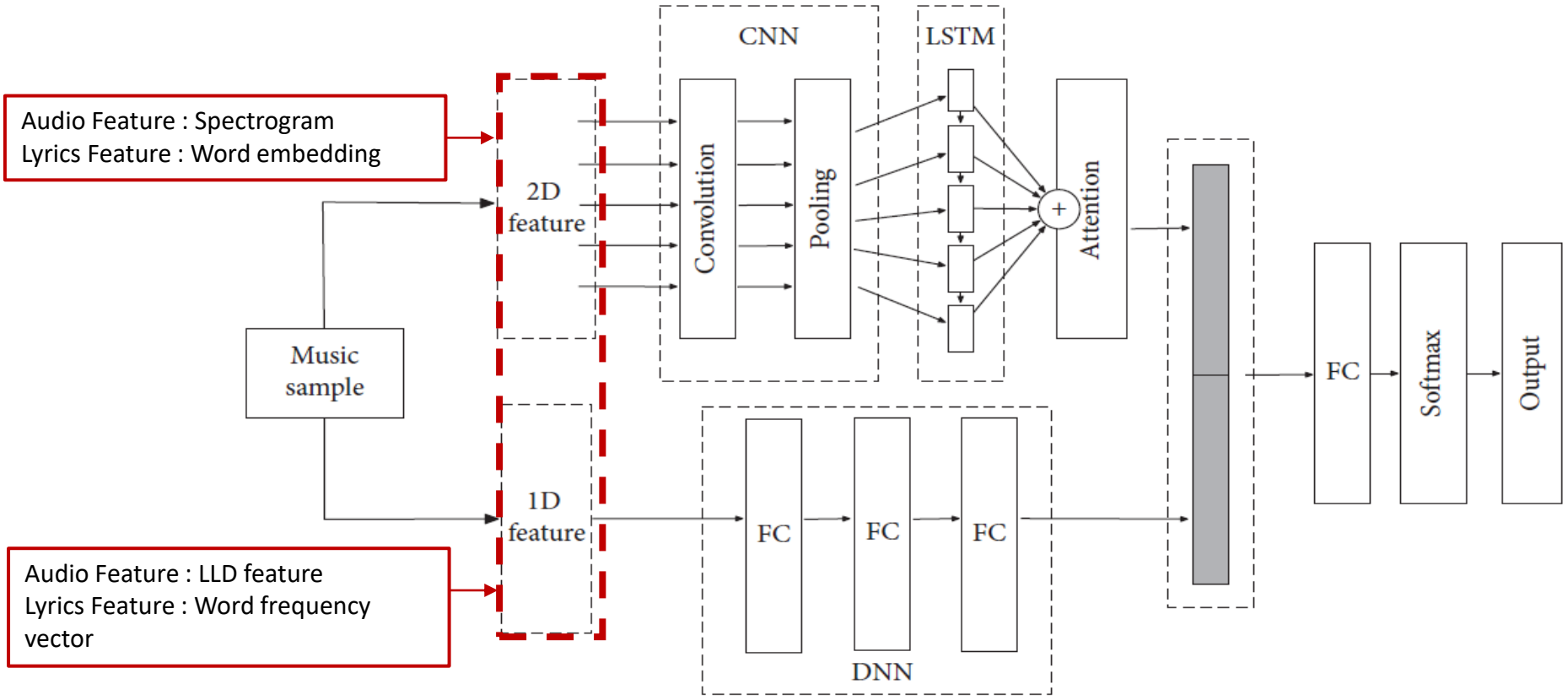




Category	Features
LLDs	Mel frequency cepstrum coefficient (MFCC), zero crossing rate (ZCR), spectral centroid, spectral spread, spectral rolloff, spectral flu, and chroma features
HSFs	Maximum, mean, and variance

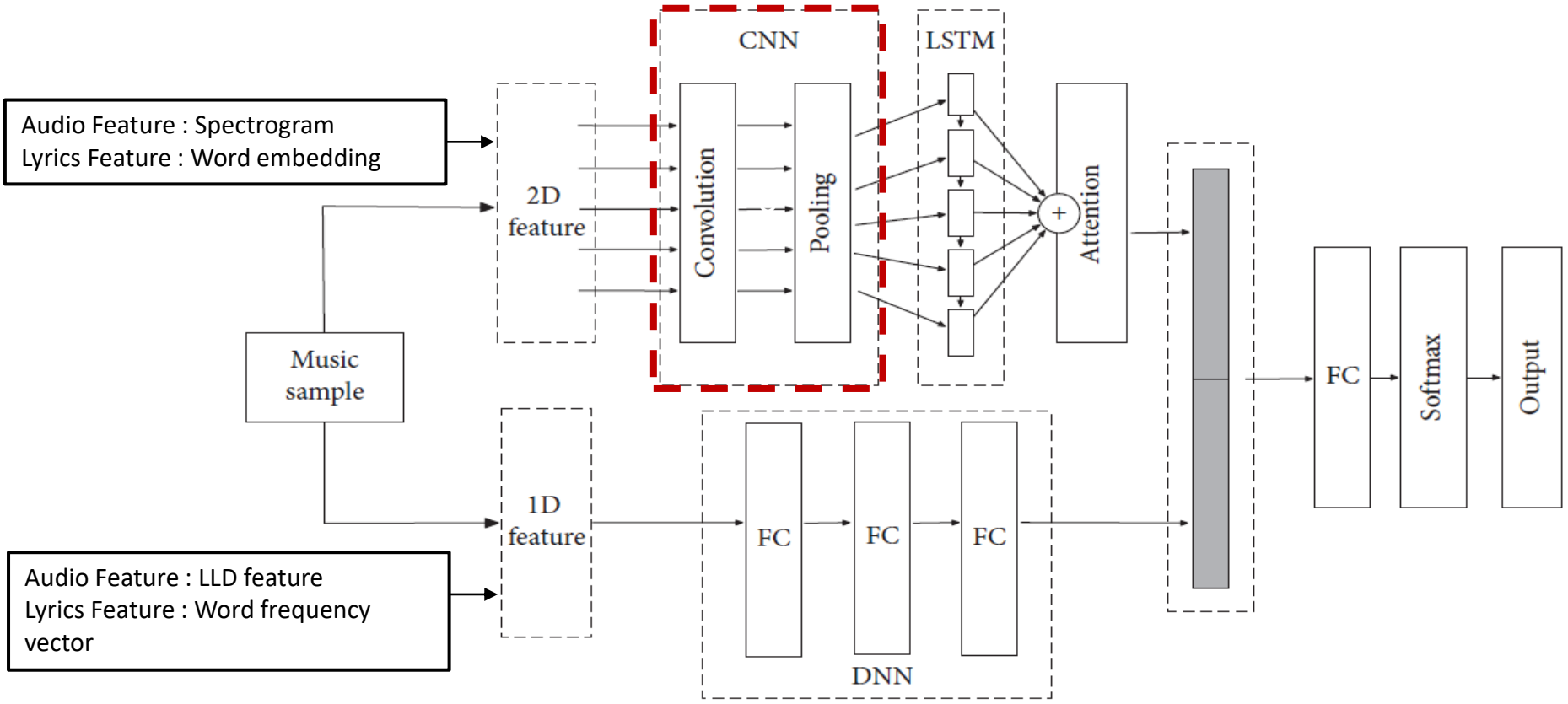
Multi feature combined network classifier based on CNN-LSTM

- Construct single-modal classifiers of audio and lyrics



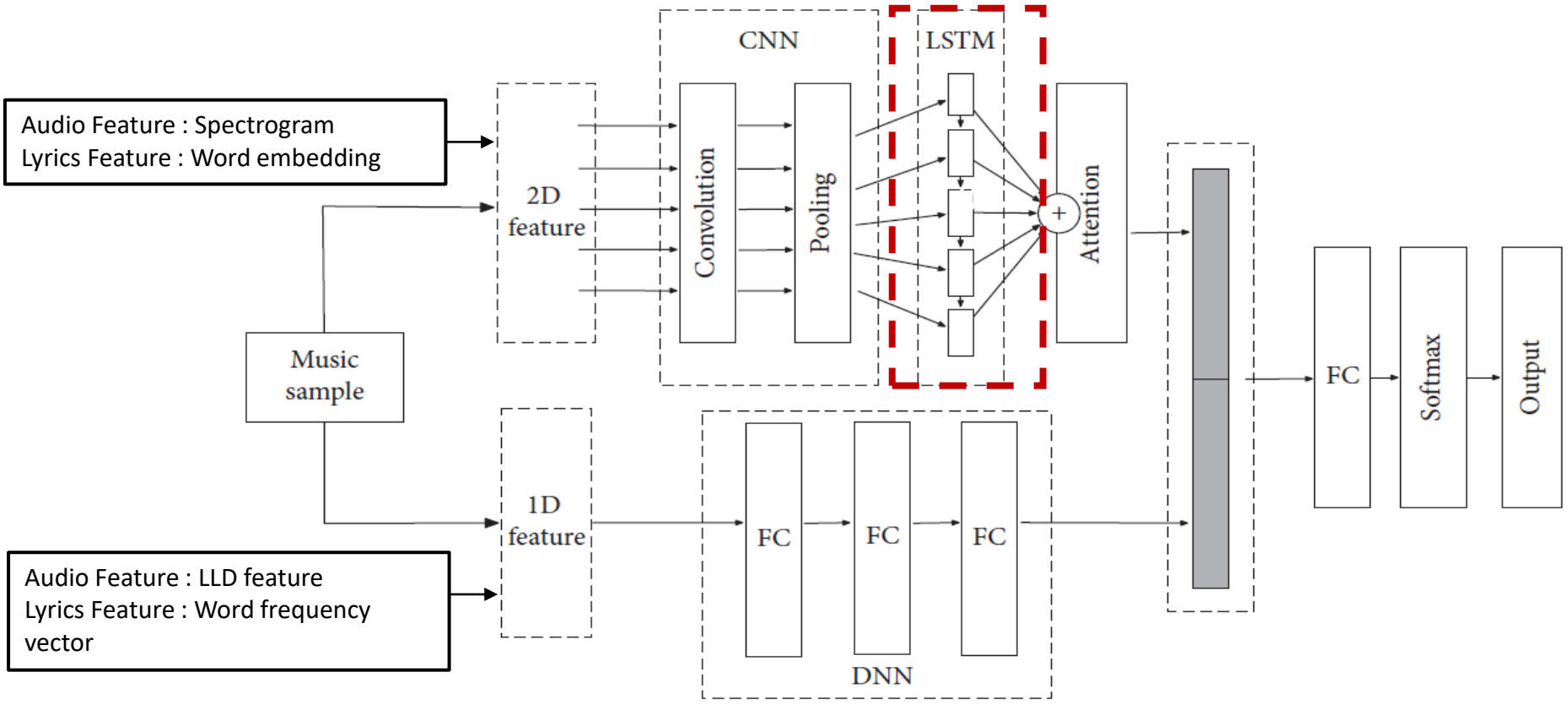
Multi feature combined network classifier based on CNN-LSTM

- The CNN layer contains 2 convolutional layers and 2 pooling layers. the first layer of convolution input is an audio spectrogram or lyrics Word2vec
- CNN is used as Feature extractor



Multi feature combined network classifier based on CNN-LSTM

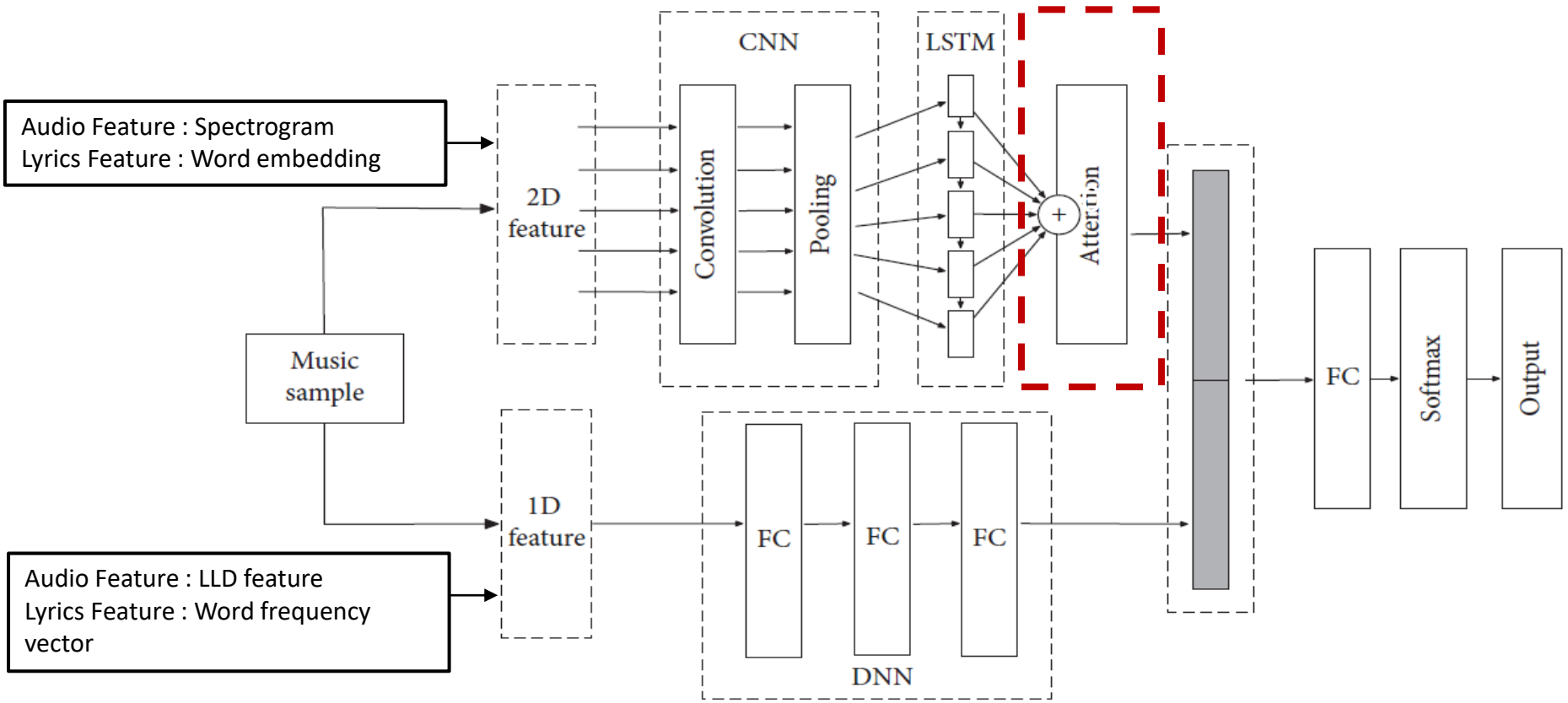
- The feature sequence output from the CNN layer can extract the features of each moment through the LSTM unit
- LSTM can effectively capture the context information of the input sequence and solve the problem of preservation and transmission of serialized information



Multi feature combined network classifier based on CNN-LSTM

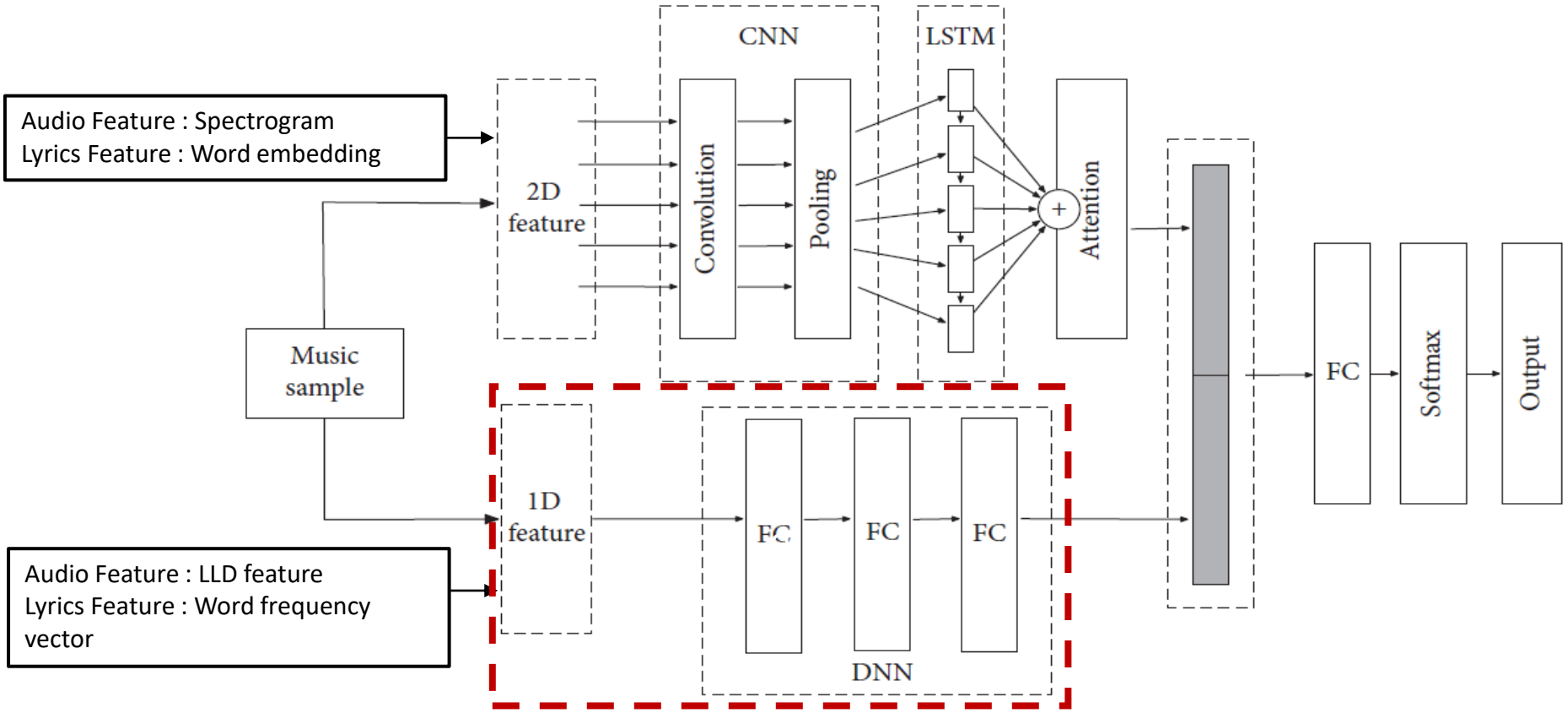
- r_i = *the number of output by the LSTM*
- a_i = *attention weight value*
- $f(r_i)$ = *score function (Softmax)*
- $\mathrm{att_n}$ = *weighted sum of the attention values of the entire sequence*

$$a_i = \frac{\exp(f(r_i))}{\sum_j \exp(f(r_j))} \rightarrow att_n = \sum_i a_i r_i$$



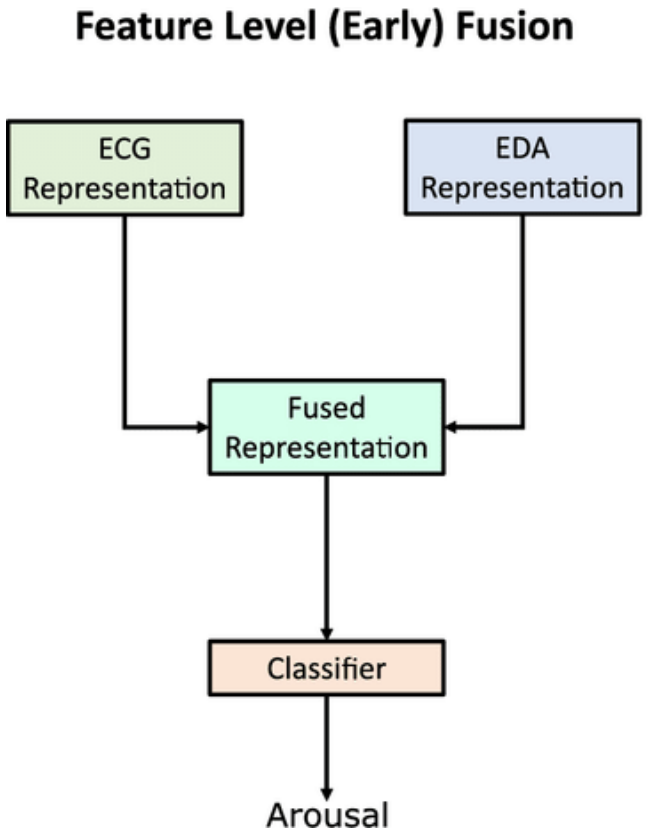
Multi feature combined network classifier based on CNN-LSTM

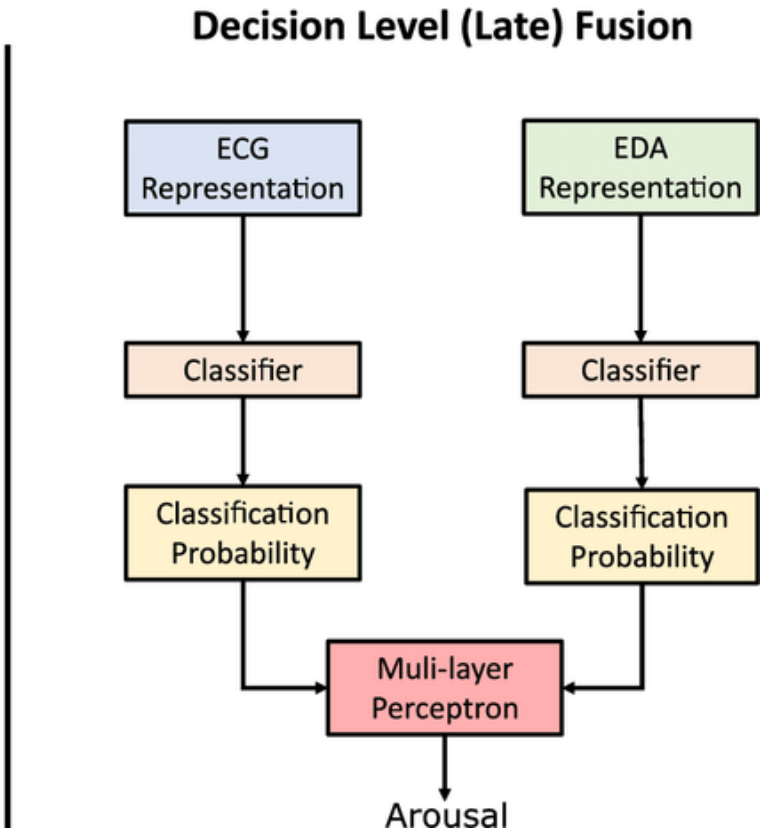
- Input 1D Feature (LLD feature, Word frequency vector)



Multimodal fusion

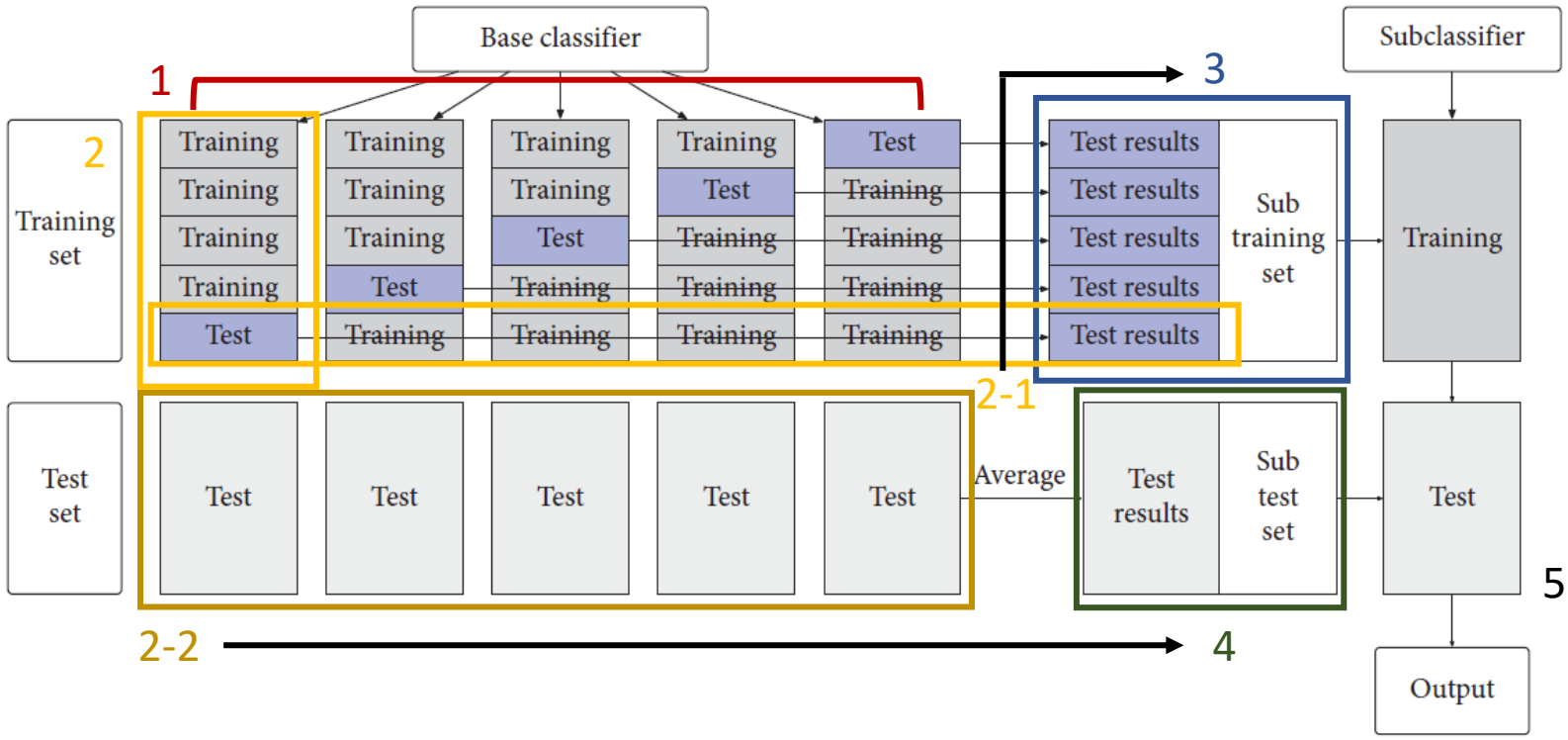
- Multimodal fusion methods in existing research generally contains two types
- Feature level fusion
- Decision level fusion





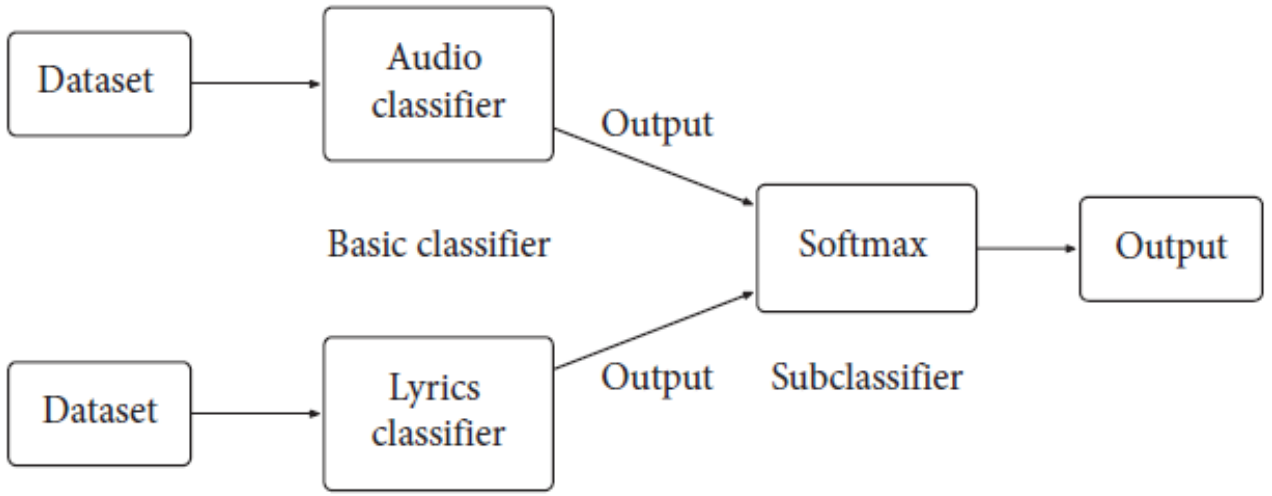
CV Stacking ensemble

- 5Fold cross-validation



Stacking ensemble

- Using softmax as subclassifier



Dataset

Audio dataset

- Last.fm tag subset of the million song dataset
- The emotional tags are angry, happy, relaxed and sad
- 500 songs are extracted from each emotional list, for a total of 2000

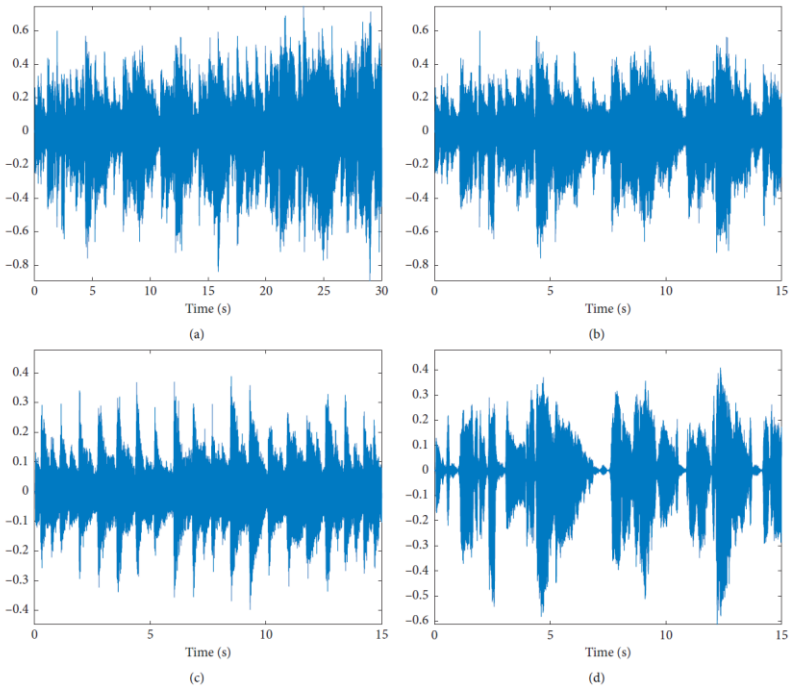
Lyrics dataset

- They used script tools to download the song audio and lyrics files in accordance with the tag lists and selected them manually

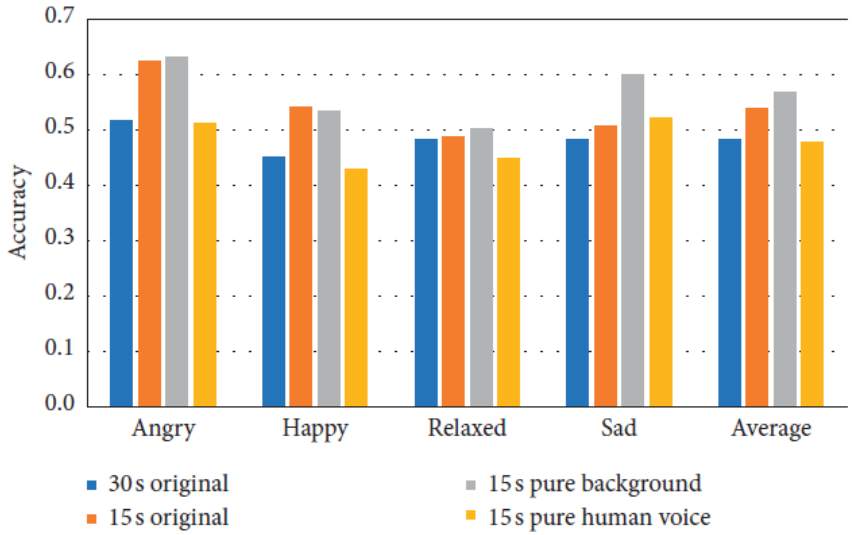
Audio waveform after preprocessing

- The audio samples are preprocessed at four levels to construct 4 experimental datasets, as shown in Figure

Audio waveform after preprocessing



The accuracy of audio classification in 4 preprocessing methods



(a) 30 s original, (b) 15 s original, (c) 15 s pure background, (d) 15 s pure human voice

Audio Experiment Result

- This group of experiments adopts different classification models to verify audio classification performance
- The multi feature combined network classifier proposed in this paper has achieved the best classification effect

Accuracy of different audio classification models

Classification models	Angry	Happy	Relaxed	Sad	Average
Spectrogram + CNN	0.643	0.594	0.51	0.62	0.592
Spectrogram + LSTM	0.631	0.427	0.54	0.438	0.509
Spectrogram + CNN-LSTM	0.632	0.632	0.632	0.632	0.632
Our model	0.705	0.602	0.688	0.73	0.681

Lyrics & Fusion method Experiment Result

- The table above is different classification models to verify lyrics classification performance
- The multi feature combined network classifier proposed in this paper has achieved the best classification effect
- The table below is different classification models to verify lyrics classification performance
- The multimodal ensemble method based on stacking proposed in this paper has obtained the best performance with an accuracy of 78%

Accuracy of different Lyrics classification models

Classification models	Angry	Happy	Relaxed	Sad	Average
Word2vec + CNN	0.584	0.62	0.615	0.671	0.622
Word2vec + LSTM	0.685	0.709	0.647	0.731	0.693
Word2vec + CNN-LSTM	0.721	0.823	0.627	0.742	0.728
Our model	0.752	0.815	0.646	0.756	0.742

Accuracy of different multimodal fusion methods

Fusion methods	Angry	Happy	Relaxed	Sad	Average
Feature fusion	0.742	0.712	0.682	0.762	0.724
Decision fusion	0.78	0.765	0.705	0.782	0.748
Our fusion method	0.808	0.823	0.726	0.773	0.782

Comparative Experiment

- The classification models proposed by other researches in the field of music sentiment classification in recent years
- The multifeatured combined network classifier proposed in this paper has better classification effects
- However, It was lower than the accuracy of 'The sentence-level decision fusion, 2017' proposed by Su, Xue

Performance comparison of proposed model with existing models

	Modal	Time	Classification	Accuracy
Seo, Huh [7]	Audio	2019	LLDs + SVM	0.571
Zhao et al. [9]	Audio	2018	MIDI + RNN	0.568
Chen, Tang [11]	Lyrics	2018	TF-IDF	0.622
Reddy, Mamidi [14]	Lyrics	2018	Word2vec + LSTM	0.693
Rachman et al. [24]	Multimodal	2018	Random forest feature fusion	0.738
Shi, Feng [25]	Multimodal	2018	LFSM decision fusion	0.758
Su, Xue [26]	Multimodal	2017	Sentence-level decision fusion	0.806
This paper	Multimodal		Multifeature combined classifier + stacking fusion	0.782

Conclusion

Conclusion

- The audio dataset was optimized through fine-grained human voice separation preprocessing, and a multi feature combined network classifier based on CNN-LSTM was proposed
- The classifier combines heterogenous 2D and 1D emotional features and has been effectively used in audio and lyrics classification, with a high classification accuracy
- Proposed a multimodal ensemble learning method based on stacking. Compared with single-modal classification, it has outstanding classification effect and remarkable generalization ability

Q & A