



WhisperMSS: A Two-Stage Framework for Mandarin Singing Transcription and Segmentation Using Pretrained Models

Ruoxuan Liang¹, Xiangjian Zeng^{2,†}, Zhen Liu³, Qingqiang Wu^{1,4,†}, RuiChen Zhang¹, Le Ren⁴

¹School of Film, Xiamen University, China

²School of Journalism and Communication, Xiamen University, China

³Zeekr Smart Network Center, Zeekr Group, China

⁴School of Informatics, Xiamen University, China

liangruoxuan@stu.xmu.edu.cn, xjzeng@xmu.edu.cn, Zhen.Liu23@zeekr.life.com,
wuqq@xmu.edu.cn, {38120241150249, renle}@stu.xmu.edu.cn

Abstract

This paper addresses the challenges of Mandarin singing transcription and segmentation by proposing a two-stage framework based on Whisper. Our research makes three key contributions: First, enhancing transcription accuracy via WhisperMLT, which incorporates a Chinese-specific text embedding layer, a CTC branch atop the encoder, and a Transformer-based contextual network; Second, optimizing CTC posterior probabilities through syllable-aligned pseudo-labeling, which generates one-hot frame-level labels from timestamp-annotated datasets; Finally, achieving precise segmentation with CTC-Vseg, which implements silence label insertion, constrained state transitions, and dynamic programming-based path optimization. Experiments demonstrate superior performance in Mandarin singing segmentation, offering novel solutions for audio processing tasks.

Index Terms: singing transcription, singing segmentation, syllable-level timestamp

1. Introduction

Fine-grained Chinese singing transcription and segmentation play a critical role in music information processing [1]. This task aims to convert unaccompanied Chinese singing audio into precise Chinese syllable/lyric sequences while identifying onset and offset time boundaries for each syllable or character. Though analogous to melody transcription, it faces heightened challenges due to the inherent flexibility and variability of vocal performances compared to instrumental audio. The outcomes of this task can further serve as foundational features for downstream applications, such as music information retrieval (MIR) and music generation, thereby enhancing their efficiency and effectiveness.

Singing segmentation can be handled by traditional rule-based or machine learning approaches, yet is constrained by dataset quality and transcription precision. Lately, training deep learning models on large weakly-supervised speech datasets has notably improved feature representation, attaining top-notch transcription and segmentation results. OpenAI, for instance, constructed a 680,000-hour weakly-supervised speech dataset from public data and trained the Whisper Seq2Seq Transformer. Employing prompt-based multi-task learning, Whisper shows outstanding zero-shot transferability, being robust and generalizable across various domains, tasks, languages, and datasets.

However, applying Whisper to Chinese singing has issues. Its multilingual label system doesn't fit Mandarin's features, causing label-character mismatches and hurting transcription accuracy. The training dataset, from auto-annotated internet sources, is imprecise. Thus, high-quality Mandarin a cappella may have transcription errors like Cantonese, traditional characters, or streaming metadata, degrading segmentation accuracy. Moreover, Whisper fails to handle common Chinese singing challenges like silent segments and elongated notes, resulting in subpar segmentation.

Traditional lyric transcription methods primarily relied on finite state automata (FSA) and signal processing techniques. For instance, Fujihara et al. employed FSA for lyric recognition [2], while Gupta et al. leveraged the repetitive structure of lyrics to enhance transcription performance [3]. Later advancements introduced signal processing-based approaches, such as cycle frequency-harmonic-time transformations (CFHT) for note-level transcription [4]. Additionally, phoneme-informed neural networks improved transcription accuracy by incorporating linguistic knowledge into the process [5].

Connectionist Temporal Classification (CTC), introduced by Graves et al. [6], has been widely adopted for sequence labeling tasks, particularly in end-to-end speech recognition [7]. Its application in lyric transcription has demonstrated effectiveness, especially when combined with weakly labeled datasets for improved accuracy [8]. CTC-based models, such as Wav2Letter [9] and deep convolutional neural networks (CNNs) [10], have further advanced transcription performance. Joint CTC-attention frameworks have been proposed to enhance sequence alignment and recognition accuracy [11][12].

Beyond transcription, CTC has also been applied to segmentation tasks, including text-to-speech alignment in long-form audio recordings [13] and large-scale corpus segmentation in speech recognition systems [14]. These studies highlight CTC's versatility in both transcription and segmentation.

The Transformer architecture [15] has significantly influenced speech processing, enabling the development of attention-based models. OpenAI's Whisper model represents a notable advancement, employing a transformer-based approach for robust multilingual transcription under diverse acoustic conditions [16]. WhisperX [17] further improves temporal accuracy in long-form audio processing [16], while variants such as WhisperT and DTW-enhanced Whisper-Crisper refine alignment precision and transcription accuracy [18][19].

Despite progress in CTC- and attention-based transcription models, challenges persist. CTC can't well capture long-range dependencies, and attention-based methods like Whisper have efficiency and alignment issues in real-time use. We propose

†Corresponding Authors

a new framework combining CTC and attention mechanisms to boost transcription and segmentation robustness.

2. Method

Our framework employs a two-stage approach for Chinese singing segmentation. First, WhisperMLT, a modified Whisper-based model, enhances Mandarin lyric transcription accuracy. Then, refined transcriptions guide segmentation using DTW and CTC-VSeg for precise temporal alignment.

2.1. WhisperMLT

Whisper faces two primary limitations in Chinese lyric transcription: (1) Its BPE-based multilingual tokenizer, trained on cross-lingual text corpora, complicates the correspondence between text units and Chinese characters, leading to many-to-one, one-to-one, or one-to-many mappings that disrupt character-level alignment. (2) The inherent annotation accuracy issues in Whisper’s cross-lingual training data can result in dialect mixing (e.g., Cantonese elements) and diverse writing forms (e.g., traditional characters) even when processing standard Mandarin singing audio.

We propose **WhisperMLT**, a transfer learning approach for Mandarin lyric transcription. It replaces Whisper’s text embedding with a Chinese-specific layer to ensure character alignment and exclude non-Chinese content. By adding a CTC branch to the encoder and a Transformer-based context network, it enhances feature extraction while preserving Whisper’s hyperparameter consistency.

2.2. WhisperMSS

For Chinese singing segmentation, CTC-based methods [14] rely on CTC loss and Viterbi [20] decoding but require large datasets and often have lower accuracy. In contrast, cross-attention enables fine-grained segmentation without extra constraints. To address low-resource challenges, we propose WhisperMSS, integrating cross-attention with syllable timestamp information for precise Mandarin singing segmentation.

The model employs multi-task learning, using Dynamic Time Warping (DTW) [21] to derive heuristic alignment paths from the cross-attention weight matrix. Additionally, CTC-VSeg extracts syllable timestamp information to guide CTC posterior probabilities, achieving syllable-level segmentation.

2.2.1. DTW and Label-Level Alignment Paths

We employ the Dynamic Time Warping (DTW) algorithm [21] to extract label-level alignment paths from cross-attention weights, which can capture monotonic alignment relationships between text labels and acoustic feature frames. These two sequences are the text label sequence $Y = \{y_1, y_2, \dots, y_L\}$ and the acoustic frame sequence $X = \{x_1, x_2, \dots, x_T\}$, where L and T denote the number of labels and frames, respectively. The cross-attention weight matrix $A \in \mathbb{R}^{L \times T}$, generated during autoregressive decoding, serves as the alignment reference. Its elements $\alpha_{i,t}$ represent the contribution of acoustic frame x_t when decoding label y_i , subject to the normalization constraint $\sum_{t=1}^T \alpha_{i,t} = 1$.

To construct the DTW cost matrix, we apply L2 normalization to the attention vector $\mathbf{A}_i = \{\alpha_{i,1}, \alpha_{i,2}, \dots, \alpha_{i,T}\}$ corresponding to each label y_i , and define the negative normalized weights as alignment costs. By constraining the path $P^* = (p_1, \dots, p_M)$ to satisfy $p_{m+1} - p_m \in \{(0, 1), (1, 1)\}$,

the DTW algorithm derives an optimal many-to-one mapping. The path endpoints $p_1 = (1, 1)$ and $p_M = (L, T)$ ensure global alignment consistency. Finally, the path P^* is converted into binary frame-level pseudo-labels, which jointly guide the optimization of the decoder’s attention distribution and the frame-level predictions of the CTC branch.

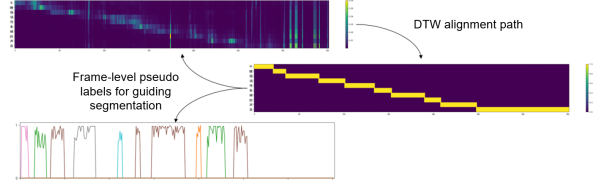


Figure 1: Label-Level Segmentation via DTW and Attention Weights

2.2.2. Frame-Level Pseudo-Label Generation for CTC

To enhance the efficiency and accuracy of the model for Chinese singing segmentation tasks, we adopt a modeling strategy based on toneless syllables as units. Specifically, 406 toneless syllable units are defined based on the *Contemporary Chinese Dictionary*, and the lyrics are converted into corresponding syllable sequences using the pinyin tool. Building upon this foundation, we perform frame-level pseudo-label generation to fully leverage labeled Chinese singing datasets with syllable-level timestamps. For a sample (X, Y) , let its syllable label sequence be $Y' = (y'_1, \dots, y'_L)$, where the onset and offset timestamps of y'_i in the dataset are $(\text{onset}_i, \text{offset}_i)$. Based on these timestamps, a frame-level label sequence $\hat{Y} \in \mathbb{R}^{|S| \times T}$ of length T is generated, where S is the syllable-based lexicon, and \hat{y}'_t represents the guidance information for the t -th frame, i.e., the one-hot encoding of the syllable label corresponding to the audio segment at that frame. This process is formally defined by Equation (2-1):

$$y_i^l = \begin{cases} \text{Onehot}(y_j) & \text{if } \lfloor \frac{\text{onset}_j}{0.02} \rfloor \leq i \leq \lfloor \frac{\text{offset}_j}{0.02} \rfloor \\ \text{Onehot}(\varepsilon) & \text{otherwise} \end{cases} \quad (2-1)$$

The intuition behind this is illustrated in an example figure. The generated frame-level label sequence \hat{Y} serves as pseudo-labels to directly guide the posterior probabilities of the CTC branch. Through this mechanism, the model learns to more precisely associate acoustic features with syllable labels, thereby improving segmentation and transcription capabilities for Chinese singing audio.

		你		好							
...	0	0	0	0	0	0	0	0	0	0	0
你	0	1	1	1	0	0	0	0	0	0	0
好	0	0	0	0	0	1	1	1	1	0	0
...	0	0	0	0	0	0	0	0	0	0	0
...	0	0	0	0	0	0	0	0	0	0	0

Figure 2: Construct fine-grained information based on timestamps

2.2.3. CTC-VSeg

In the task of Chinese singing segmentation, traditional CTC-Segmentation algorithms fail to effectively handle potential silent segments between labels, resulting in suboptimal segmentation. To accurately extract label timestamps, this paper proposes an improved segmentation method based on the Viterbi algorithm—CTC-VSeg. This method treats the transcription label sequence as an observation sequence, uses CTC posterior probabilities as state transition probabilities, and searches for the globally optimal segmentation result through the Viterbi algorithm [20]. Specifically, let the lyric content be $Y = (y_1, \dots, y_L)$ and the syllable sequence be $Y' = (y'_1, \dots, y'_L)$. The CTC branch of the model predicts a frame-based posterior probability matrix $P \in \mathbb{R}^{T \times (|S|+1)}$, where $p_{t,j}$ represents the probability of the transcription label y_j at time step t . To handle silent segments, the algorithm inserts silent labels ε at the boundaries and between labels, resulting in an augmented lyric sequence \tilde{Y} of length $2L + 1$. Based on the posterior probability matrix, the algorithm computes the cumulative score matrix $A \in \mathbb{R}^{T \times (2L+1)}$ through dynamic programming, where $a_{t,j}$ represents the maximum joint probability of aligning the sequence $\tilde{Y}[1 : j]$ with the audio frames $X[1 : t]$. The algorithm designs three valid state transition rules: (1) transitioning from the $(i-1)$ -th label to the i -th label, indicating no silent segment between them; (2) transitioning between non-silent labels and adjacent silent labels; and (3) maintaining the current label or silent segment, indicating its continuation over time. The cumulative score matrix A is computed using dynamic programming according to Equation (2-2):

$$\begin{cases} a_{t-1,j} + \log p_{t,\varepsilon} & \text{if } j = 0 \\ \max(a_{t-1,j}, a_{t-1,j-1}, a_{t-1,j-2}) + \log p_{t,j} & \text{if } \tilde{y}_j \neq \tilde{y}_{j-2} \\ \max(a_{t-1,j}, a_{t-1,j-1}) + \log p_{t,j} & \text{otherwise} \end{cases} \quad (2-2)$$

Additionally, the matrix B records the state transition paths, where the element $b_{t,j}$ indicates the previous path step when computing $a_{t,j}$. The final segmentation result is determined by backtracking the path $\arg \max_{t,j} (a_{t=T,j=2L+1}, a_{t=T,j=2L})$.

2.2.4. Multi-Task Loss Function

We propose a multi-task loss function to optimize the Chinese singing voice segmentation task, comprising four key component losses. First, the autoregressive decoder loss \mathcal{L}_{s2s} measures the discrepancy between the predicted label sequence from the AED branch and the ground-truth labels via negative log-likelihood, thereby enhancing label prediction accuracy. Second, the guided cross-attention weights loss \mathcal{L}_{GAW} leverages a binarized matrix $A \in \mathbb{R}^{L \times T}$ to generate frame-level pseudo-labels, constraining the attention weights to focus on critical acoustic segments and improving the model's ability to capture salient acoustic information. Third, the CTC branch guidance loss \mathcal{L}_{GCTC} aligns CTC output probabilities with expected distributions using the matrix

$$M_{CTC} = \sum_{i=1}^L y_i A_i^T \quad (2-3)$$

and the loss term

$$-\sum (M_{CTC} \odot \log P_{CTC}). \quad (2-4)$$

Finally, to address the "spike phenomenon" in singing segmentation, we introduce a maximum entropy-regularized CTC loss (EnCTC):

$$\mathcal{L}_{EnCTC} = \mathcal{L}_{CTC} - \beta H(p(\pi|Y, X)), \quad (2-5)$$

where β controls the weight of the entropy regularization term. By incorporating the conditional entropy of alignment paths, EnCTC suppresses excessive spikes in CTC outputs and prevents overfitting to blank-dominated paths, promoting more reasonable segmentation. The overall loss is defined as:

$$\mathcal{L} = \mathcal{L}_{EnCTC} + \mathcal{L}_{s2s} + \lambda (\mathcal{L}_{GAW} + \mathcal{L}_{GCTC}),$$

where the hyperparameter λ balances the influence of hard-alignment guidance. Joint optimization of these losses strengthens the model's capability to learn fine-grained acoustic segmentation patterns, ultimately improving singing voice segmentation performance.

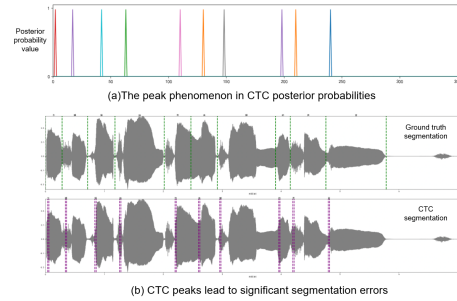


Figure 3: CTC peak phenomenon

3. Training

3.1. Datasets

We use two Chinese lyric transcription datasets: M4Singer [22] and Opencpop [23]. M4Singer, from Zhejiang University and Huawei, contains 20,896 samples (30 hours) of multi-singer Mandarin singing with sheet music, lyrics, and timestamps, covering 700 pop songs. Opencpop, by the Wenet team, includes 3,756 samples (5.2 hours) from professional female singers, with syllable and phoneme pitch labels and boundary timestamps.

3.2. WhisperMLT

WhisperMLT adopts a two-stage transfer learning process for Chinese lyric transcription. In the first stage, the Encoder and Decoder layers of the model are frozen, and the CTC branch and the pure Chinese text embedding layer are fine-tuned using the AISHELL1 Mandarin speech dataset [24] for two epochs. The model is then fully fine-tuned on all layers using the M4Singer Chinese singing dataset. This stage unfreezes all layers, allowing the model to update parameters based on the characteristics of singing audio. The best-performing model is then tested on the full data of Opencpop to compare transcription performance across different models. During training, M4Singer is split into a 90% training set and 10% validation set, with Levenshtein edit distance used to ensure no overlap between the validation and training samples.

3.3. WhisperMSS

This section adopts the same data partitioning strategy as the previous section, utilizing the full dataset of M4Singer for training and the full dataset of Opencpop for testing. For the hyper-parameters in Equations (3-1) and (3-10), we first fix $\beta = 1$ and determine the optimal value of λ within the range $[0, 1]$ with a step size of 0.2. Subsequently, we fix λ to its optimal value and apply the same strategy to search for the optimal β . The model achieves the best performance on the validation set with $\lambda = 0.2$ and $\beta = 0.4$.

We adopt Average Absolute Error (AAE) to evaluate segmentation accuracy, computing it separately for syllable onset and offset timestamps. Given the significant duration variability in singing compared to conversations, separate AAE metrics offer a comprehensive assessment of boundary segmentation precision.

3.4. Implementation Details

The experiment was carried out on a high performance server running the Ubuntu 22.04.3 LTS operating system. The hardware setup includes an NVIDIA A800 80GB PCIe GPU.

4. Evaluation

4.1. Metrics

For the lyric transcription task, we evaluate performance using Character Error Rate (CER) and Syllable Error Rate (SyER). CER is computed as the Levenshtein edit distance between predicted and true character sequences, normalized by the true sequence length. SyER is derived by converting characters to toneless syllables using the pypinyin tool and computing the error rate. Both metrics are defined as:

$$\text{CER/SyER} = \frac{S + D + I}{L} \quad (4-1)$$

where S , D , and I represent substitution, deletion, and insertion errors, respectively, and L is the length of the true sequence.

For the singing segmentation task, we use Average Absolute Error (AAE) to measure timestamp accuracy. Let L be the lyric length, and let onset'_i , offset'_i , onset_i , and offset_i denote the predicted and true start/end times (in seconds) for the i -th label. AAE is calculated as:

$$\text{AAE} = \frac{1}{2L} \sum_{i=1}^L (|\text{onset}'_i - \text{onset}_i| + |\text{offset}'_i - \text{offset}_i|) \quad (4-2)$$

Additionally, we adopt the Percentage of Correct Onsets (PCO) metric, which considers a prediction correct if the timestamp error is within a tolerance of 0.02 seconds. Precision (P), recall (R), and F1-score (F1) are computed for both start and end times to evaluate segmentation accuracy from a classification perspective.

4.2. Results

4.2.1. Syllable-Level Mandarin Lyrics Transcription

We evaluated the Conformer model and two self-supervised models, Wav2vec2.0 [25] and HuBERT [26]. Conformer was fine-tuned on AISHELL1, then M4Singer, and tested on Opencpop. Wav2vec2.0 and HuBERT used Facebook AI’s pre-trained checkpoints, fine-tuned on M4Singer. OpenAI’s pre-trained model was tested zero-shot for accuracy assessment.

Model	Decoding Method	CER	SyER
Conformer _{finetuned}	beam-search	11.4	4.9
wav2vec2 _{large} +CTC	greedy-search	7.0	2.2
HuBERT _{large} +CTC	greedy-search	6.7	2.1
Whisper Pretrained _{small}	beam-search	13.2	5.4
Whisper Pretrained _{medium}	beam-search	10.7	2.8
Ours/WhisperMLT _{small}	beam-search	5.1	1.3
Ours/WhisperMLT _{medium}	beam-search	4.6	1.1

Table 1: Performance Comparison of Speech Recognition Models

Table 3.3 shows our method outperforms in Chinese lyric transcription, with the fine-tuned WhisperMLT-medium model achieving the lowest CER of 4.6% and the lowest SyER of 1.1%. Converting characters to syllables significantly reduces errors across models, indicating frequent phonetic accuracy with semantic misinterpretations. Notably, Whisper_{small} and Whisper_{medium} achieve SyERs of 1.3% and 1.1%, respectively. Compared to other methods, WhisperMLT excels in capturing pronunciation details in singing audio, enhancing its potential for singing evaluation tasks.

4.2.2. Syllable-Level Mandarin Lyrics Segmentation

We select the wav2vec 2.0 Fully - Connected Layer method (weakly - supervised speech segmentation based on pre - trained models) and Jun’s 2023 Whisper [27] - based Mandarin transcription method as baselines. Table [specific number] shows model performance. The proposed WhisperMSS method uses fine - grained timestamps to guide CTC branch frame - level multi - classification, with syllable segmentation AAE as low as 0.02 s and onset F1 of 92.3% (0.02 s tolerance). All models perform better on onset than offset due to variable singing pronunciation unit durations. Ablation experiments were conducted to assess the proposed multi - information fusion method. Removing DTW - based label - level segmentation extraction increases AAE from 0.020 s to 0.1 s, and reduces F1_{onset} and F1_{offset}. Removing CTC - VSeg also increases AAE and slightly reduces F1_{onset} and F1_{offset}. These results show the value of each fused information in improving model performance, validating the method’s effectiveness and superiority.

Model	AAE(s)	F1 _{onset}	F1 _{offset}
wav2vec2-fc	0.150	90.4	90.1
Jun	0.022	91.9	91.8
WhisperMSS	0.020	92.3	92.1
(w/o)DTW	0.100	90.5	90.2
(w/o)CTC-VSeg	0.023	92.0	91.7

Table 2: Performance Comparison of Speech Segmentation Models

5. Conclusion

We proposed WhisperMSS, enhancing Mandarin singing transcription and segmentation by refining the Whisper model with a CTC branch. Frame-level pseudo-labeling and CTC-Vseg methods enable precise syllable segmentation. Future work will explore its potential in speech defect quantification and cross-language applications.

6. References

- [1] B. Bhattarai and J. Lee, “A comprehensive review on music transcription,” *Applied Sciences*, vol. 13, no. 21, p. 11882, 2023.
- [2] T. Hosoya, M. Suzuki, A. Ito, S. Makino, L. A. Smith, D. Bainbridge, and I. H. Witten, “Lyrics recognition from a singing voice based on finite state automaton for music information retrieval.” in *ISMIR*, 2005, pp. 532–535.
- [3] M. McVicar, D. P. Ellis, and M. Goto, “Leveraging repetition for improved automatic lyric transcription in popular music,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 3117–3121.
- [4] Y. Wu, Y. Ju, S. Lui, J. Yang, F. Fan, and X. Du, “Cycle frequency-harmonic-time transformer for note-level singing voice transcription,” in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.
- [5] S. Yong, L. Su, and J. Nam, “A phoneme-informed neural network model for note-level singing transcription,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [6] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.
- [7] A. Graves and N. Jaitly, “Towards end-to-end speech recognition with recurrent neural networks,” in *International conference on machine learning*. PMLR, 2014, pp. 1764–1772.
- [8] Y. Qiu, J. Zhang, Y. Shan, and J. Zhou, “Enhancing note-level singing transcription model with unlabeled and weakly labeled data,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 341–345.
- [9] R. Collobert, C. Puhersch, and G. Synnaeve, “Wav2letter: an end-to-end convnet-based speech recognition system,” *arXiv preprint arXiv:1609.03193*, 2016.
- [10] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C. L. Y. Bengio, and A. Courville, “Towards end-to-end speech recognition with deep convolutional neural networks,” *arXiv preprint arXiv:1701.02720*, 2017.
- [11] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, “Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm,” *arXiv preprint arXiv:1706.02737*, 2017.
- [12] S. Kim, T. Hori, and S. Watanabe, “Joint ctc-attention based end-to-end speech recognition using multi-task learning,” in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4835–4839.
- [13] G. Doras, Y. Teytaut, and A. Roebel, “A linear memory ctc-based algorithm for text-to-voice alignment of very long audio recordings,” *Applied Sciences*, vol. 13, no. 3, p. 1854, 2023.
- [14] L. Kürzinger, D. Winkelbauer, L. Li, T. Watzel, and G. Rigoll, “Ctc-segmentation of large corpora for german end-to-end speech recognition,” in *International Conference on Speech and Computer*. Springer, 2020, pp. 267–278.
- [15] A. Waswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *NIPS*, 2017.
- [16] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.
- [17] M. Bain, J. Huh, T. Han, and A. Zisserman, “Whisperx: Time-accurate speech transcription of long-form audio,” *arXiv preprint arXiv:2303.00747*, 2023.
- [18] R. Wang, Z. Xu, and F. X. Lin, “Efficient whisper on streaming speech,” *arXiv preprint arXiv:2412.11272*, 2024.
- [19] L. Wagner, B. Thallinger, and M. Zusag, “Crisperwhisper: Accurate timestamps on verbatim speech transcriptions,” *arXiv preprint arXiv:2408.16589*, 2024.
- [20] G. D. Forney, “The viterbi algorithm,” *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 2005.
- [21] T. Giorgino, “Computing and visualizing dynamic time warping alignments in r: the dtw package,” *Journal of statistical Software*, vol. 31, pp. 1–24, 2009.
- [22] L. Zhang, R. Li, S. Wang, L. Deng, J. Liu, Y. Ren, J. He, R. Huang, J. Zhu, X. Chen *et al.*, “M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 6914–6926, 2022.
- [23] Y. Wang, X. Wang, P. Zhu, J. Wu, H. Li, H. Xue, Y. Zhang, L. Xie, and M. Bi, “Opencpop: A high-quality open source chinese popular song corpus for singing voice synthesis,” *arXiv preprint arXiv:2201.07429*, 2022.
- [24] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, “Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline,” in *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.
- [25] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
- [26] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.
- [27] J.-Y. Wang, C.-I. Leong, Y.-C. Lin, L. Su, and J.-S. R. Jang, “Adapting pretrained speech model for mandarin lyrics transcription and alignment,” in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.