MM-ALT: A Multimodal Automatic Lyric Transcription System

Xiangming Gu*
Integrative Sciences and Engineering Programme, NUS
Graduate School, National University of Singapore
Singapore
xiangming@comp.nus.edu.sg

Longshen Ou*
School of Computing, National University of Singapore
Singapore
longshen@comp.nus.edu.sg

Danielle Ong
School of Computing, National University of Singapore
Singapore
danielleong@u.nus.edu

Ye Wang
School of Computing, National University of Singapore
Singapore
wangye@comp.nus.edu.sg

ABSTRACT

Automatic lyric transcription (ALT) is a nascent field of study attracting increasing interest from both the speech and music information retrieval communities, given its significant application potential. However, ALT with audio data alone is a notoriously difficult task due to instrumental accompaniment and musical constraints resulting in degradation of both the phonetic cues and the intelligibility of sung lyrics. To tackle this challenge, we propose the MultiModal Automatic Lyric Transcription system (MM-ALT), together with a new dataset, N20EM, which consists of audio recordings, videos of lip movements, and inertial measurement unit (IMU) data of an earbud worn by the performing singer. We first adapt the wav2vec 2.0 framework from automatic speech recognition (ASR) to the ALT task. We then propose a video-based ALT method and an IMU-based voice activity detection (VAD) method. In addition, we put forward the Residual Cross Attention (RCA) mechanism to fuse data from the three modalities (i.e., audio, video, and IMU). Experiments show the effectiveness of our proposed MM-ALT system, especially in terms of noise robustness. Project page is at https://n20em.github.io.

CCS CONCEPTS

• **Applied computing** → **Sound and music computing**; • **Information systems** → **Music retrieval**; **Speech / audio search**; • **Computing methodologies** → *Neural networks*; • **Human-centered computing** → Ubiquitous and mobile computing systems and tools.

KEYWORDS

lyric transcription, multimodality, neural networks, dataset

---

*Both authors contributed equally to this research.

1 INTRODUCTION

Automatic lyric transcription (ALT) is the task of recognizing lyric text from singing. The supplementing of textual information via ALT often facilitates solutions to other music information retrieval problems, such as lyric alignment [28], query by singing [32], audio indexing [26], and music subtitling [25]. Additionally, ALT systems can be augmented to appraise singing from various linguistic or musical aspects, such as word intelligibility and pronunciation[58]. Recently, research in ALT has become increasingly active, as evidenced by the emergence of benchmark singing datasets, e.g., [18, 55, 56], and further studies of advanced acoustic modeling techniques, e.g., [9, 21].

To our knowledge, ALT systems thus far are built with audio-only data, which has arguably led to a plateau in research efforts plagued by the following reoccurring challenges:

Difficulty of the ALT task. Since musicality distinguishes singing from speech, a singer inevitably sacrifices some richness of linguistic features such as word stress and articulation to compensate for musical features such as melody, tempo, and timbre. Consequently, singing is generally less intelligible than speech [69]. This insufficient richness of linguistic features handicaps approaches adapted from automatic speech recognition (ASR) to ALT. This might account for the performance gap (19.60% vs. 8.7% word error rate) of the same acoustic model between ALT [18] and a similar-scale ASR task [64].

Nonconformity with human perception patterns. The perception of speech extends beyond the auditory realm, as observed by the McGurk effect [51] that visual information can substantially contribute to auditory perception. Studies have shown that visual cues of speech play an essential role in language learning [19, 53], lending further support to the notion that human auditory perception relies on multimodal information.

Insufficient robustness. Given the two points above, it is reasonable to surmise that attempting ALT on audio-only recordings in challenging signal-to-noise ratio (SNR) environments would probably yield unsatisfactory results. Coupled with lowered intelligibility over speech and the lack of external cues, the added factor of noise (e.g., accompaniments) further increases the difficulty of retrieving information from musical audio.

Due to the interconnected relationship between sound production and observable lip movements, it is clear that adding the modality of visual information to the ALT system is a promising prospect at present. Recent advances in audio-visual speech recognition (AVSR), e.g., [50, 71], show that audio-visual models perform better than their audio-only counterparts, suggesting that including additional sensory information on top of the acoustic signal enhances the overall performance of ALT systems by enriching the input to the model. We assume that training an ALT system on multimodal data is a viable solution to the challenges mentioned earlier.

Additionally, inertial measurement unit (IMU) data from wearable devices are an augmentation offering potential for improving system performance. In the early stages of data collection, we observed that participants tended to make rhythmic head movements and facial expressions along with the music, in addition to regular pronunciation efforts during singing. This was especially obvious for those who displayed higher levels of musicality. Furthermore, the jaw movements of singers are highly correlated with the pronunciation of lyrics. We assume that these additional cues are valuable in improving lyric recognition and increasing system robustness. We use *eSense* [41], an earbud augmented with inertial sensors, to capture data of facial and head movements. This device has been proven to reliably detect head-, jaw-, and mouth-related movements [49, 65, 67], facial expression [45], speakers in conversations [57], and even human emotions [65].

We conduct this work to validate our assumptions. The contributions are summarized into three aspects:

- We present the MultiModal Automatic Lyric Transcription system (MM-ALT), which utilizes three modalities of input: audio, video, and signals from wearable IMU sensors. To facilitate building the system, we curate the N20EM dataset for multimodal lyric transcription. We create a group of models on this dataset that can perform multimodal lyric transcription with varying combinations of modalities, obtaining a minimum word error rate (WER) of 12.71%. We further reveal an increase in system robustness by introducing additional modalities. With severe perturbations of musical accompaniments (-10 dB SNR), our system can achieve 27.04% absolute lower WER compared to its audio-only counterparts.
- We initialize two new tasks: lyric lipreading and IMU-based voice activity detection (VAD). In the lyric lipreading task, we attempt to recognize lyrics in singing utilizing only video information. Our video encoder is the first attempt to retrieve language-related information from singing recordings without the help of audio input. As to the IMU-VAD task, our IMU encoder is the first attempt at building a frame-level VAD system solely from motion data captured by a wearable IMU device. Our experiments elucidate the correlation between the IMU and audio modalities.
- We propose Residual Cross Attention (RCA), a new feature fusion method to better fuse the multimodal features using self-attention and cross-attention mechanisms. We demonstrate the effectiveness of this new modality fusion method in our ALT system by comparing it with various feature fusion methods.

2 RELATED WORK

2.1 Automatic Speech Recognition

The development of deep neural network (DNN) techniques has revolutionized ASR by prompting a shift in paradigm from hidden-Markov-model- (HMM-) based systems, to DNN-HMM hybrid models, and finally, to end-to-end (E2E) models [46]. Compared to previous models, E2E ASR systems possess more consistent optimization objectives, more simplified pipelines, and more compact model structures[46]. The Connectionist Temporal Classification (CTC) [27] and Attention-based Encoder-Decoder (AED) [6, 14] are two mainstream E2E ASR models. Jointly optimizing the shared encoder of AED and CTC models in a multitask learning framework greatly improves convergence and mitigates the misalignment issue of AED models [31, 42].

Training state-of-the-art ASR models under supervised learning framework requires a large amount of transcribed data, which is too demanding for low-resource languages. In recent years, it has become a trend to use self-supervised learning (SSL) techniques to build ASR systems. By utilizing easy-to-access unlabeled data, SSL models can learn general data representations to facilitate downstream tasks. Models such as wav2vec 2.0 [5], autoregressive predictive coding [17], and HuBERT [33] can learn very powerful semantic representations of audio. These models can achieve impressive results even with very limited labeled data, demonstrating the feasibility of low-resource speech recognition.

It is convenient to incorporate multimodal input into E2E ASR models, especially a simultaneous audio and video signal. Audio-visual speech recognition (AVSR) brings machines closer to how humans perceive speech. Lots of research supports the notion that adding video input to the ASR systems has positive effects on recognition performance, e.g., [50, 63, 71, 73, 83]. In particular, Audio-Visual Hidden Unit BERT (AV-HuBERT) [70, 71] notably outperforms previous state-of-the-art models in both lipreading and AVSR tasks by learning the joint representation of a synchronized speech and audio signal.

2.2 Automatic Lyric Transcription

ALT is the counterpart problem of ASR in the field of music information retrieval. The research of ALT started from [32], by adapting an HMM model to build a Japanese lyric transcription system. Then, Mesaros and Virtanen studied the influence of in-domain lyric language models (LM) [54]. [52] leveraged musical structure to solve the ALT problem by using the repetitive chorus of one song to improve the consistency and accuracy of their ALT system. Afterward, Kawai et al. first attempted a deep learning approach to ALT by a DNN-HMM model [40]. In the baseline system constructed for the DSing dataset [18], a contemporary state-of-the-art acoustic model TDNN-F was used for ALT. Demirel et al. then utilized a convolutional neural network (CNN) and time-restricted self-attention to extract more robust features before feeding the input to the TDNN-F [20]. Gupta et al. tried building a genre-informed acoustic model for ALT systems directly on polyphonic singing audio [28]. Demirel et al. also built a multi-stream TDNN-F model [21] that performed better than the original version. More recent network architectures such as Transformers have also been used in ALT systems [9].

One reason for the slower progress in developing ALT systems compared to advancements in ASR, is the inaccessibility of large-scale singing datasets [81]. Therefore, some researchers have focused on providing basic support for the ALT problem from a data standpoint. For example, Dabike et al. built a pipeline to preprocess the DAMP Sing! 300x30x2 dataset [11] to form an utterance-level dataset suitable for ALT [18], and Meseguer-Brocal et al. built a large-scale popular music dataset, DALI, containing synchronized audio, lyrics, and notes [55, 56].

Another approach to alleviate the problem of insufficient data is data augmentation. For instance, Kruspe adjusted speech data to make it more "song-like" by random time stretching and pitch adjusting [44]. Basak et al. used a vocoder-based synthesizer to generate singing voices according to natural speech [9]. Lastly, Zhang et al. proposed aligning lyrics with melodies before adjusting duration and pitch during data augmentation [81].

2.3 Multimodal Learning

Humans perceive the world via different modalities, *e.g.*, vision, sound, touch, and smell. These modalities have heterogeneous and complementary information, enabling humans to understand their surroundings better. Therefore, it is natural to design multimodal systems to process and fuse multiple modalities' high-dimensional inputs simultaneously. Given recent advancements in deep learning techniques, multimodal methods have surpassed their single-modal counterparts in empirical applications. For example, [34, 68] combined RGB and depth images to improve semantic segmentation; [1, 50, 70, 71, 74] adopted video and audio modalities to boost the performances of ASR and active speaker detection systems; [35, 75] fused multi-view images or videos and IMU data to benefit 3D human pose estimation. Apart from empirical results in previous literature, Huang et al. proceeded to formalize the multimodal learning problem into a theoretical framework. They proved that learning with multiple modalities results in more accurate estimated representations of latent spaces, which can achieve better performance of learning than using its subset of modalities [36].

One fundamental problem in multimodal learning is to integrate multimodal data in a way that can exploit the complementary relationship and redundancy of multiple modalities [8]. A common technique is to fuse individual modality inputs or intermediate features together. This fusion operation could be concatenation, e.g., [50, 70, 75, 77], addition, e.g., [34, 35, 68], or incorporation of an attention mechanism, e.g., [13, 74, 79, 80, 82]. Given the success of transformers in computer vision [23, 47], natural language processing [22, 76] and speech processing [5, 33], transformers are now widely used to learn multimodal representations. Among them, Perceiver models iteratively distill multimodal inputs into a tight latent bottleneck and build latent transformers to process these latent features [10, 38, 39]. Nagrani et al. proposed the framework of a Multimodal Bottleneck Transformer (MBT), which fuses the multimodal data via attention bottlenecks [59].

3 DATASET

This section describes the construction process and the preprocessing procedures of our N20EM dataset.

3.1 Corpus Curation

We adopted the same 20 songs used in [24] for practical reasons of rich phonemic coverage, ease of learning, and variation in musical features like genre and tempo. Instead of excising repeated lyrics and scat singing, we opted to preserve all utterances for completeness of data and naturalness of singing from participants. Seventeen males and thirteen females were recruited from a local university, with musical backgrounds ranging from no formal vocal training to amateur level exposure. Participants spoke English with accents including North American, European, Indian, East Asian, and Southeast Asian. Participants were free to choose songs they were more familiar and comfortable with singing from the 20 songs. However, we limited each song only to be selected by a maximum of 10 participants to encourage a greater selection diversity within each participant's repertoire and achieve a more balanced dataset.

3.2 Data Collection Procedure

All data were recorded in a soundproof studio. Our setup consisted of three main components:

- An Audio-Technica 4050 condenser microphone with a pop filter was used to record audio data at 32-bit depth and 44.1 kHz sampling rate using Adobe Audition. A monaural headset was used for playback of musical accompaniment to participants during singing.

- An earbud, *eSense* [41], containing several built-in components, including an IMU sensor, a speaker, and a microphone, was used to record the kinesthetic activity of the head, jaw, and lips during recording at a sampling rate of 100 Hz.

- A Sony AX4 video camera and a ring light were placed in front of participants, focused on the lower half of each singer's face, to capture footage of movements of oral articulators (jaw, lips, tongue) at 1920x1080 pixels, at 50 Hz frame rate.

Participants were instructed to minimize bodily movements such as hand gestures and body swaying to reduce noise factors in data, such as accidental obstruction of video footage or anomalous variance in the sensory data. Lyric sheets for all songs were printed and placed on a music stand by the microphone for the subject's reference. The tempo and key of each song were predetermined. For comfort in vocal range, participants were given the option of a male-vocals arrangement or female-vocals arrangement for songs with greater variance in melodic range. Although the backing tracks fed to participants included vocals, participants were given some freedom in the rendition of pitch and rhythm during singing. When a track with all the lyrics clearly sung was obtained, the subject proceeded to the next song. A few pronunciation errors were allowed as long as the utterance remained clear.

Since singing is often studied in tandem with speech, we also collected read versions of the song lyrics for future comparative studies. For each song, the selected lyrics were read first, then sung with the aid of a backing track in a separate recording afterward. We believe this is beneficial for future research on the quantitative difference between singing and parallel speech[1].

---

[1]Presently, the parallel speech recordings are not used in the study since the focus of this paper is to build an ALT system.

Table 1: Division of data into different sets for ALT system

Set	Duration	Number of Utterances
Total	5 h 23 min	5116
Train	4 h 1 min	3803 (74%)
Validation	35 min	616 (12%)
Test	47 min	697 (16%)

3.3 Data Preprocessing

Data from different modalities were first synchronized with 40 ms precision. For video recordings, following the work of [70], we first down-sampled videos to 25 Hz, then cut out 96×96 Regions-Of-Interest (ROIs) centered around the mouth. Only cropped videos containing lip areas will appear in the public dataset to protect participant privacy. Audio data were down-sampled to 16 kHz to match the input specification of wav2vec 2.0 [5].

We opted for an utterance-level annotation of lyrics. First, we manually annotated the starting and ending timestamps of each utterance. The standard of annotation was determined mainly by natural factors such as musical cadence or practical aspects such as preferring consonant boundaries over vowel boundaries between utterances. The annotation was done using spectrogram information and the marker function in Adobe Audition. Second, actually sung words were transcribed to serve as the ground truth for ALT. When they were different from the correct lyrics that should be sung, we used a set of notations to mark different types of errors. We believe this information will be helpful for future singing pronunciation evaluation research on this dataset[2]. Please refer to Appendix A for the annotation details.

After obtaining the annotation, the recordings of different modalities were all segmented to utterance level according to the annotation. If any, silence, breaths, or non-phonemic noise in between utterances were excised from the data. Then, the dataset was split into training, validation, and test sets. We ensure that no subset contained utterances from the same song to test the generalization abilities of the ALT systems. Table 1 shows the statistics of different subsets.

4 MM-ALT SYSTEM

In this section, we first formalize the task of multimodal ALT and then describe our proposed MM-ALT for addressing the problem.

4.1 Problem Formulation

We consider the multimodal setting for automatic lyric transcription. More specifically, given the synchronized singing audio signal $\mathbf{x}^{(a)}$, video signal $\mathbf{x}^{(v)}$, and IMU signal $\mathbf{x}^{(i)}$, our goal is to transcribe the lyrics, i.e., obtain a word sequence $w_{1:S}$ representing the corresponding lyrics of the signal. As shown in Fig. 1, we propose the MM-ALT system to solve this problem. The system G consists of a feature representation learning frontend and automatic lyric transcription backend. Firstly, modality-specific encoders $G_{\text{enc}}^{(a)}, G_{\text{enc}}^{(v)}, G_{\text{enc}}^{(i)}$ are adopted to extract the features for each modality of signal. Then the feature fusion module G_{fuse} projects the features from different

---
[2]The error annotations are not utilized in the current system.

modalities into the same latent space and integrates them into more representative features. Finally, the hybrid CTC-Attention backend G_{back} transforms the sequence of fused features into the lyrics. The whole system can be represented as:

$$
\begin{aligned}
w_{1:S} &= G(\mathbf{x}^{(a)}, \mathbf{x}^{(v)}, \mathbf{x}^{(i)}) \\
&= G_{\text{back}}(G_{\text{fuse}}(G_{\text{enc}}^{(a)}(\mathbf{x}^{(a)}), G_{\text{enc}}^{(v)}(\mathbf{x}^{(v)}), G_{\text{enc}}^{(i)}(\mathbf{x}^{(i)})))
\end{aligned} \quad (1)
$$

4.2 Audio Encoder

The audio encoder $G_{\text{enc}}^{(a)}$ aims to learn acoustic representations from the audio signal. Traditionally, the TDNN network and its variants dominate the field of ALT [18, 20, 21]. In this paper, we propose to utilize wav2vec 2.0 [5] as the audio encoder for ALT through a transfer learning paradigm because wav2vec 2.0 generalizes well into new domains with low-resource labeled data. From our experiments, we notice that when accepting only the audio signal, our system achieves state-of-the-art performance, exceeding the results of all published approaches on the DSing dataset [18], one of the mainstream lyric transcription datasets.

Wav2vec 2.0 consists of a feature encoder and a context network. The feature encoder has seven blocks, each containing a temporal convolution followed by a layer normalization and a GELU activation. It takes raw audio and outputs latent speech representations. The context network contains 24 transformer blocks with model dimension 1,024, inner dimension 4,096, and 16 attention heads. It transforms latent speech representations into context representations by capturing global temporal information. Each frame of final output $\mathbf{f}^{(a)}$ is about 20 ms and has 1,024 dimensions. We refer readers to Appendix B for the detailed implementations. To transfer the knowledge of wav2vec 2.0 from the speech domain into the singing domain, we remove the quantization module in the original wav2vec 2.0 structure and fine-tune the model on singing datasets.

4.3 Video Encoder

The video encoder $G_{\text{enc}}^{(v)}$ seeks to learn visual representations of speech from cropped lip videos. Since this is the first attempt to transcribe lyrics from video modality, there are currently no benchmark models. We propose adopting the Audio-Visual Hidden Unit BERT (AV-HuBERT) [70] in our system, which is the state-of-the-art approach in the task of lip reading for speech recognition.

AV-HuBERT consists of an image encoder and a backbone transformer encoder. The image encoder is built by a 3D convolutional layer followed by a ResNet-18 block [29]. Similar to the context network in wav2vec 2.0, the transformer encoder has 24 blocks, each of which has a model dimension of 1,024, a feed-forward dimension of 4,096, and 16 attention heads. Each frame of resulting representations $\mathbf{f}^{(v)}$ is about 40 ms and has 1,024 dimensions. We remove the audio layer in the original structure and only feed the video modality into AV-HuBERT.

4.4 IMU Encoder

To validate the assumption that IMU motion data correlates with the audio signal, and further exploit such correlation to help the lyric transcription, the IMU encoder $G_{\text{enc}}^{(i)}$ converts IMU signal to features that correspond to speaking. We utilize a convolutional-recurrent neural network (CRNN) containing 1D convolutional layers and
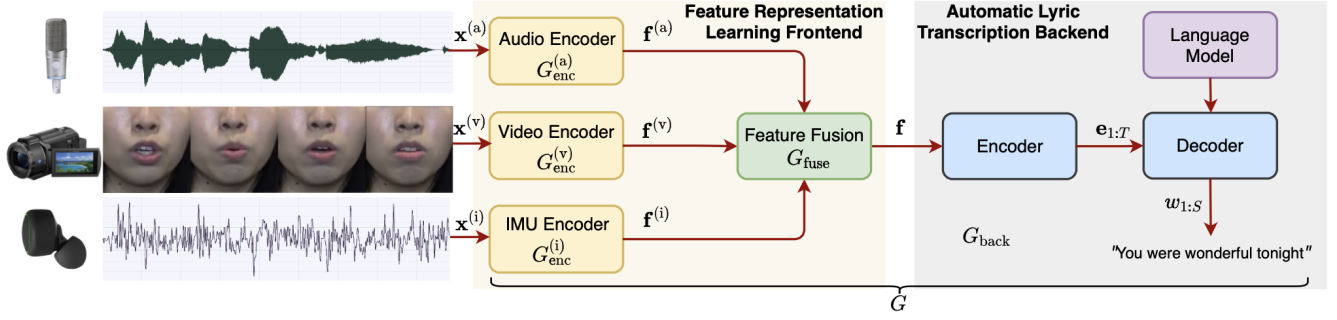
Figure 1: An overview of the MM-ALT system.

Figure 2: Structure of IMU CRNN.

bi-directional Gated Recurrent Unit (GRU) layers, resembling the best-performing network CNN-BiLSTM in [60]. Fig. 2 shows the structure of our IMU CRNN. The input of CRNN contains 8 channels, including three axes of accelerometer data, three axes of gyroscope data, and the quadratic sums of different channels of the two sensors, respectively. When incorporated into the ALT system, the output of the last GRU layer $\mathbf{f}^{(i)}$ is used as the IMU features. Each $\mathbf{f}^{(i)}$ has 20 ms frame length and 120 dimensions.

4.5 Feature Fusion Module

The feature fusion module G_{fuse} is designed to exploit the complementary relationship and redundancy of different modalities. As illustrated in Fig. 3, the dimensions of features from different modalities are firstly unified. We up-sample the video to the same time resolution as the features of audio and IMU. Then we pad or truncate the features of the other two modalities to standardize the number of frames amongst the three modalities before the fusion operation. For IMU features, we adopt a linear layer following a GELU activation to increase its dimensions from 120 to 1024.

We propose a new attention module named Residual Cross Attention (RCA). RCA is built upon Transformer block architecture. An RCA block accepts inputs from multiple modalities. One input is regarded as the source, providing keys and values, while other inputs are considered as the reference, which provides queries. Besides the multi-head self-attention (MHSA) [76] operation on the source, RCA adds extra shortcuts by computing the multi-head cross-attention (MHCA) operation between the source and each reference. RCA can be represented by Eq. 2 and Eq. 3:

$$\mathbf{f}' = \text{LN}(\mathbf{f}_{\text{src}} + \text{MHSA}(\mathbf{f}_{\text{src}}) + \text{MHCA}(\mathbf{f}_{\text{src}}, \mathbf{f}_{\text{ref1}}) + \text{MHCA}(\mathbf{f}_{\text{src}}, \mathbf{f}_{\text{ref2}})) \quad (2)$$

$$\mathbf{f} = \text{LN}(\text{FFN}(\mathbf{f}') + \mathbf{f}') \quad (3)$$

where LN refers to a layer normalization layer, and FFN is a positional-wise feed forward network the same as the one in a Transformer block [76]. There are three RCA modules during the feature fusion, and each modality's input serves as the source in one of the modules. Finally, we add the output together and obtain the final fused features $\mathbf{f}$.

The motivations behind the proposal of RCA in multimodal scenarios are explained as follows. Firstly, RCA adopts self-attention and residual shortcuts to extract global relationships among features of all frames and reserves information of source modality. Secondly, RCA takes advantage of complementary information from reference modalities through its cross-attention mechanism. Specifically, the queries from reference modalities help downstream modules delve into missing relationships between time frames that are not attended to when using only self-attention.

4.6 Hybrid CTC-Attention Backend

Inspired by [78], we design a hybrid CTC-Attention backend to map a sequence of fused features $\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_T \in \mathbb{R}^{1024}$ into a sequence of tokens $w_1, w_2, ..., w_S \in \mathbb{V}$, where $\mathbb{V}$ is the vocabulary. In our implementations, $\mathbb{V}$ has 31 tokens including 30 character targets and a word boundary token.

Firstly, we adopt a 2-layer MLP as the encoder to further encode the fused representations $\mathbf{f}_{1:T}$ into $\mathbf{e}_{1:T} \in \mathbb{R}^{T \times 1024}$. Then we design two branch networks, one of which is used to output CTC
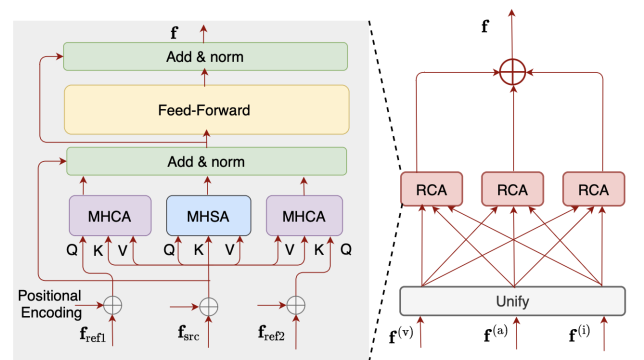

Figure 3: An illustration of the feature fusion module and the proposed Residual Cross Attention (RCA) algorithm.

predictions and another is used as the decoder to output Sequence-to-Sequence (S2S) predictions. Specifically, a projection layer is used to map $\mathbf{e}_{1:T}$ into the output probabilities for each frame $p(\pi_t|\mathbf{e}_t)$, where the CTC predictions $\pi_t \in \mathbb{V}$. $\mathbf{e}_{1:T}$ are also fed into the S2S decoder, which is parameterized via a location-aware attention-based GRU [15], to be autoregressively decoded into target probabilities $p(w_c|w_{1:c-1}, \mathbf{e}_{1:T})$. $p(\pi_t|\mathbf{e}_t)$ and $p(w_c|w_{1:c-1}, \mathbf{e}_{1:T})$ are used for two different training and inference schemes, which will be elaborated in the next section.

4.7 Training and Inference

The whole system is trained with cross-entropy loss of outputs from both connectionist temporal classification (CTC) [27] and sequence-to-sequence (S2S) [7] attention-based decoder. Suppose the ground truth transcription is $w_1, w_2, ..., w_S$. The CTC loss can then be written as:

$$\mathcal{L}_{\text{CTC}} = -\log p_{\text{CTC}}(w_{1:S}|\mathbf{e}_{1:T})$$
$$= -\log \sum_{\pi_{1:T} \in \mathcal{B}^{-1}(w_{1:S})} \prod_{t=1}^{T} p(\pi_t|\mathbf{e}_t) \quad (4)$$

Here, $\mathcal{B}$ is an operation mapping an alignment sequence $\pi_{1:T}$ to $w_{1:S}$ by removing repeated tokens and word boundary tokens. $\mathcal{B}^{-1}(w_{1:S})$ refers to all the CTC paths mapped from $w_{1:S}$.

The S2S loss can be written as:

$$\mathcal{L}_{\text{S2S}} = -\log p_{\text{S2S}}(w_{1:S}|\mathbf{e}_{1:T})$$
$$= -\log \prod_{c=1}^{S} p(w_c|w_{1:c-1}, \mathbf{e}_{1:T}) \quad (5)$$

The overall loss function is the addition of the two loss terms above. We introduce a hyper-parameter, α, to balance both losses.

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{\text{S2S}} + \alpha\mathcal{L}_{\text{CTC}} \quad (6)$$

During inference, the most likely lyrics will be predicted considering the output of CTC, S2S, and language model (LM), by the following equation:

$$w_{1:S}^* = \arg\max_{w_{1:S}} \beta \log p_{\text{CTC}}(w_{1:S}|\mathbf{e}_{1:T})$$
$$+ (1-\beta) \log p_{\text{S2S}}(w_{1:S}|\mathbf{e}_{1:T}) + \gamma \log p_{\text{LM}}(w_{1:S}) \quad (7)$$

Here, β and γ are two hyper-parameters used to balance three log-probability terms during the beam search. When $\gamma > 0$, the LSTM language model is enabled. Beam size is set as 512.

5 EXPERIMENTS

In this section, we evaluate our proposed MM-ALT system using curated N20EM dataset. Specifically, we firstly conduct single-modality experiments to evaluate the modality-specific representation learning. Then we evaluate the whole system in multimodal scenarios to demonstrate its effectiveness. Finally, we simulate the realistic environments by adding musical accompaniments as perturbations to test the robustness of our MM-ALT system.

5.1 Implementation Details

We build our MM-ALT system using the PyTorch library and Speech-Brain toolkit [66]. As for more detailed model configurations, please refer to Appendix B. We apply data augmentation during training:

Table 2: WER(%) of ALT systems on Singing datasets. We compare different methods on DSing dataset and build benchmarks for N20EM dataset. "w. DSing" refers to adding DSing to the training data.

Method	Dataset	Validation	Test
TDNN-F [18]	DSing	23.33	19.60
CTDNN-SA [20]	DSing	17.70	14.96
MSTRE-Net [21]	DSing	-	15.38
Ours	Dsing	**13.26**	**14.56**
Ours	N20EM	12.74	19.68
Ours w. DSing	N20EM	**9.65**	**13.00**

for the audio signal, we perform SpecAugment [62] in the time domain; for the video signal, we randomly flip and crop face images with the size of 88 following [70]. All models are trained using the Adam optimizer. For the ALT task, we report the Word Error Rate (WER) as evaluation metrics.

5.2 Single-Modality Tasks

5.2.1 Automatic Lyric Transcription. We evaluate the performance of audio encoder together with the automatic lyric transcription backend on the curated N20EM dataset and DSing [18] dataset, one of the mainstream ALT datasets. Firstly, wav2vec 2.0 [5] is pretrained on LibriVox (LV-60k) and loaded into our audio encoder[3]. Then we train the models on training split, validate/test the ALT performance on validation/test split. In inference, we use CTC-S2S-LM to decode the lyrics. For the DSing dataset, we train an LSTM LM only on text corpus from DSing training split, which is a subset of text corpus used by other baselines. While for the N20EM dataset, we train an LSTM LM on text corpus from LibriSpeech [61], DSing [18] and the N20EM dataset. We keep this LM configurations for other experiments on the N20EM dataset. Besides, we enable the training of the fusion module and set $\mathbf{f}^{(v)}$ and $\mathbf{f}^{(i)}$ as zero tensors for a fair comparison with the multimodal settings.

The evaluation results are summarized in Table 2. We observe that our proposed system achieves 13.26% and 14.56% WER on the DSing dataset, outperforming TDNN and its variants [18, 20, 21] by at least 4.44% on the validation split and 0.40% on the test split. The performance of our system indicates that we can successfully adapt the pre-trained model from the speech domain to the singing domain by exploiting the similarities between speech and singing.

Since our proposed ALT system displays great capability on the DSing dataset, we adopt it as a strong baseline for the N20EM dataset. To begin with, we train the system using only the N20EM dataset and observe 12.74% WER on the validation split and 19.68% WER on the test split. When adding the DSing dataset to the training data, performance is improved to 9.65% and 13.00% WER on validation and test splits respectively, demonstrating that having more singing data during the training enhances the ability of the system to generalize.

5.2.2 Lyric Lipreading. We initialize a new task in this section, named as lyric lipreading. This task aims to recognize lyrics only

---
[3]https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self

Table 3: WER(%) of lyric lipreading on N20EM dataset. We compare different decoding configurations.

CTC	S2S	LM	Validation	Test
√	×	×	63.52	78.20
×	√	×	55.72	74.10
√	√	×	55.80	72.70
√	√	√	**47.91**	**68.45**

Table 4: Results of frame-level classification for VAD using IMU signal.

Metric (%)	Validation	Test
Accuracy	69.63	74.36
F1-score	79.72	82.86
Macro F1-score	62.71	67.09

through video modality. Since this is the very first attempt, we train our video encoder together with automatic lyric transcription backend to build a benchmark system. Firstly, AV-HuBERT [70] is pretrained on LRS3 [2] and VoxCeleb2 [16], and then loaded into our video encoder[4]. Afterwards, we train the models on the N20EM dataset. Unlike the original implementation of unigram tokenizer in [70], we adopt a character tokenizer as mentioned in Section 4.6 to facilitate feature fusion and comparison among different modalities.

During experiments, we evaluate different decoding configurations. As shown in Table 3, we observe that the performance of the S2S model exceeds the CTC model by a large margin. Furthermore, adopting a trained language model can also enhance system performance. Our best model achieves 47.91% WER on the validation split and 68.45% WER on the test split.

5.2.3 Voice Activity Detection from IMU. From our experiments, it is difficult to obtain a functional IMU feature extractor by directly training the IMU encoder on the ALT task, because IMU signal is not as informative as acoustic data when the goal is to predict words. To maximize the use of IMU data, rather than immediately tackling the problem of ALT with only IMU data, we first attempt a relevant task of lower complexity: voice activity detection (VAD), the task of detecting zones of vocal activity. VAD with audio as inputs is usually an essential step to preprocess data for ASR systems to lower the downstream processing latency [3, 12]. Inspired by this, we hope to provide additional meaningful inductive bias for consequent transcription modules by performing VAD training on the IMU encoder. We integrate the IMU encoder into our multimodal ALT system after it is well-trained on the VAD task.

The annotation of VAD is obtained from audio recordings. For each recording, we select a loudness threshold that is slightly higher than the loudness of the sound of breaths. We get a rough VAD annotation by treating all frames with a volume above this threshold as voiced and the rest as unvoiced. Afterward, IMU signals of songs are segmented into 5-second clips for training.

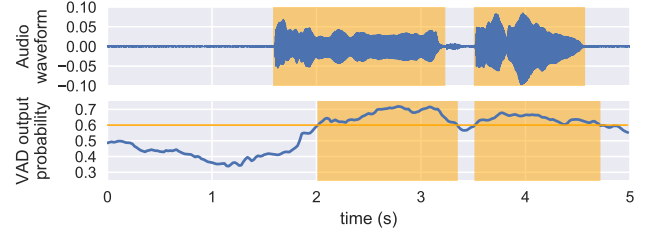[4]https://dl.fbaipublicfiles.com/avhubert/model/lrs3_vox/vsr/self_large_vox_433h.pt

Figure 4: An audio clip and the corresponding IMU VAD's output. Highlighted area represents actual/predicted voiced segments. The orange horizontal line in the VAD output figure is the classification threshold.

Table 4 shows the results of our IMU VAD model. On the test split, we achieve an accuracy of 74.36%, an F1 score of 82.86%, and a 67.09% macro F1. These results suggest the correlation between the wearable IMU sensor and microphone recording data, and our model can successfully capture this correlation. Fig. 4 shows the relationship between audio waveform and IMU VAD output in a more intuitive way. We can find that VAD output probabilities are low at unvoiced frames while becoming higher when audio frames contain voice activities.

5.3 Multi-Modality ALT

We evaluate our multimodal ALT system in Fig. 1 on the N20EM dataset. Before training the whole system, we adopt a transfer learning strategy by loading the modality-specific encoders trained in single-modality experiments. Unlike in section 5.2, we also simulate realistic environments by adding corresponding musical accompaniments to the audio signal with different signal-to-noise ratios (SNRs). We train and evaluate our system on the mixed data above to test system robustness. We compare different modality combinations in $-10, -5, 0, 5, 10$ dB SNRs, as well as clean scenarios (no accompaniments).

5.3.1 Quantitative Analysis. Quantitative results are reported in Table 5. Firstly, we compare the audio-only system with the audio-visual system. We notice that the audio-visual system performs much better than audio-only in low SNR scenarios. For example, in -10 dB, the audio-visual system significantly outperforms the

Table 5: WER(%) of various modality combinations in different SNR scenarios on N20EM dataset. A: Audio-only ALT, A-V: Audio-Visual ALT, A-V-I: Audio-Visual-IMU ALT.

SNR (dB)	Validation			Test		
	A	A-V	A-V-I	A	A-V	A-V-I
-10	74.33	42.50	**41.87**	90.96	64.83	**63.92**
-5	45.45	36.91	**36.91**	63.02	54.68	**54.68**
0	25.28	**22.70**	22.81	35.70	**34.42**	34.44
5	16.06	15.57	**15.06**	**22.19**	22.38	22.89
10	**12.37**	12.96	12.51	17.41	17.41	**17.03**
∞ (clean)	9.65	9.56	**9.45**	13.00	12.81	**12.71**
Avg.	30.52	23.37	**23.10**	40.38	34.42	**34.27**

Table 6: Qualitative Results. Insertions are marked in <u>underline and red</u>, and substitutions are marked in *italics and cyan*.

		Text
Clean	Reference	Son of god love's pure light
	A	<u>The</u> son of god *loves* pure *life*
	A-V	Son of god *loves* pure *life*
	A-V-I	Son of god *loves* pure *ligh*
Noisy	Reference	Wonder how I got along
	A	*What* is how I got *a* lot
	A-V	*Wander* how I got *a* long
	A-V-I	Wonder how I got *a* <u>long</u>

audio-only system by about 30% WER. As the SNR level increases, the performance of the two systems gradually converges. When there are no accompaniments, an audio-visual system achieves 9.56% WER on the validation split and 12.81% WER on the test split, which exceeds the performance of an audio-only system. We also report the average WER on the above six scenarios and find that the audio-visual system outperforms the audio-only system by 7.15% and 5.96% on the validation and test splits respectively, demonstrating its superior robustness.

After adding the IMU modality, we notice that our 3-modal MM-ALT system achieves the best results in the majority of scenarios. The MM-ALT system achieves 23.10% and 34.27% WER on the validation and the test splits on average, surpassing both audio-only and audio-visual configurations. Without accompaniments, our multimodal system achieves 9.45% WER on the validation split and 12.71% WER on the test split, which is the best result obtained for the N20EM dataset. Admittedly, the improvements brought by IMU modality is not significant, which indicates that IMU modality is not as promising as video modality. However, its effectiveness is still not negligible. We leave it for future work to maximize the effectiveness of the IMU modality.

5.3.2 Qualitative Analysis. Apart from the quantitative analysis, we also show the qualitative results in Table 6. We show one case from the clean scenario and one from a noisy (mixed with accompaniments) scenario. More quantitative results are displayed in Appendix C. In the clean case, the audio-only system has three word errors, including one insertion and two substitutions. Both audio-visual and audio-visual-IMU models misspell "love's" and "light", but for the word "light" the three-modal system's output has fewer character-level errors (only one missing character). Likewise, in the noisy case, the MM-ALT system also performs better than its audio-only and audio-visual counterparts by correcting the substitutions of "What" and "Wander" as well as the insertion of "is". Although the word "along" in reference is not fully recovered, the transcription from the MM-ALT system is closer to the ground truth than the transcription from the audio-only system.

5.3.3 Ablation Study. To validate the effectiveness of our proposed RCA mechanism, we conduct an ablation study for the feature fusion module in our MM-ALT system. The results are summarized in Table 7. To magnify the differences, we evaluate the ALT performance in -10 dB SNR scenario. We find that without cross attention

Table 7: Ablation study of the Residual Cross Attention (RCA) in -10dB scenario of three modalities. CA: Cross Attention, SA: Self Attention.

Method	Validation	Test
w/o CA	42.47	66.93
w/o SA	42.67	64.47
RCA	**41.87**	**63.92**
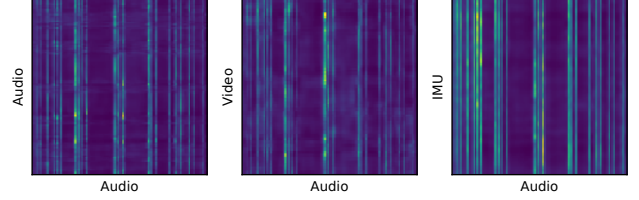
Figure 5: Visualization of self-attention and cross-attention weights in RCA module of audio modality. We use brighter colors to highlight stronger attention.

shortcuts, the ALT performance will drop 0.6% and 3.01% WER on the validation and test splits. Without a self-attention mechanism, the ALT performance will decrease by 0.8% and 0.55% WER respectively. These results suggest that RCA contributes to better feature fusion.

To further prove the effectiveness of the RCA mechanism, we visualize the attention maps in the RCA module when audio modality serves as the source, including self-attention and cross attention maps, as shown in Fig. 5. We observe that these three attention maps have captured common attention patterns. Furthermore, audio-video cross attention and audio-IMU cross attention can also extract missing relationships between time frames that are not captured by self-attention.

6 CONCLUSION

In this paper, we proposed a multimodal lyric transcription system, MM-ALT. Firstly, we curated a multimodal ALT dataset N20EM to facilitate building the system. Then, we presented an audio-based transcription method that surpasses existing state-of-the-art approaches. Next, we built modality-specific encoders based on video and IMU modalities. For video, we attempted the new task of lyric lipreading. For the IMU modality, we made the first attempt at frame-level VAD, which showed promising results and demonstrated the correlation between IMU and audio data. Finally, we used RCA, our proposed feature fusion method, to fuse information from different modalities to obtain the final transcription. Our series of attempts revealed that introducing additional modalities improves transcription performance and makes the model more robust to perturbations of musical instruments.

ACKNOWLEDGMENTS

A ANNOTATE PRONUNCIATION ERRORS IN DATASET

The paradigm used for resolving linguistic errors in utterances is illustrated in Table 1. We use different brackets to represent different error types. At the phoneme level, mispronounced words were marked with front slashes when they were found to be poorly enunciated or simply wrong. Simple substitutions were noted by including the original word in square brackets following its substitute. Inserted words were marked with curly brackets in their utterances, and deleted words were put in parentheses.

Table 8: Annotation for differnet types of pronunciation error.

Error	Example	Resolution
Mispronunciation	"little" - "laytle"	/little/
Substitution	"the" - "a"	a [the]
Insertion	"" - "a"	{a}
Deletion	"and" - ""	(and)

B IMPLEMENTATION DETAILS

B.1 wav2vec 2.0

We have included a brief description of the wav2vec 2.0 structure in Section 4.2 Audio Encoder. It consists of a feature encoder and a context network [5]. The feature encoder is a multi-layer convolutional neural network consisting of seven blocks. Each block has a 1D temporal convolution followed by a layer normalization [4] and a GELU activation [30]. The 1D temporal convolutions have 512 channels, {10,3,3,3,3,2,2} kernel sizes and {5,2,2,2,2,2,2} strides for each block. Suppose the input to the feature encoder is $\mathbf{x}$, which is 1D raw audio; its output $\mathbf{z}$ is a 2D latent representation of the input $\mathbf{x}$. The frequency of $\mathbf{z}$ is 49 Hz, and each frame has 1,024 dimensions. After that, $\mathbf{z}$ are fed into the context network to obtain the final output $\mathbf{f}$, which has been explained in Section 4.2. In addition to these two parts, wav2vec 2.0 also has a quantization module for unsupervised pretraining. Through this quantization module, latent speech representation $\mathbf{z}$ are discretized to $\mathbf{q}$, which serve as the targets for unsupervised pretraining. Since we don't need the pretraining stage, we remove the quantization module in the original structure in our experiments.

B.2 AV-HuBERT

We have included a concise description of the AV-HuBERT structure in Section 4.3 Video Encoder. The original structure is able to accept both audio and video modalities [70]. Since we only need AV-HuBERT to extract the visual representations, we remove layers handling audio inputs. The video clips are firstly fed into a 3D convolutional frontend. The 3D convolutional frontend has a 3D convolution with $5 \times 7 \times 7$ kernel size and $1 \times 2 \times 2$ stride, a batch normalization [37], a ReLU activation and a 3D max pooling with $1 \times 3 \times 3$ kernel size and $1 \times 2 \times 2$ stride. Then a modified ResNet-18 [29] is adopted to extract the latent visual features. To compensate for the absence of audio modality, we concatenate the latent visual features with zero tensors and remove the modality dropout in the original structure. The fused features are fed into the context

Table 9: More Qualitative Results. Deletions are marked in (brackets and red), and substitutions are marked in *italics and cyan*.

SNR		Text
∞ (Clean)	Ref.	Goodbye papa please pray for me
	A	Goodbye *pop* please pray for me
	A-V	Goodbye *bap* please pray for me
	A-V-I	Goodbye papa please pray for me
10 dB	Ref.	Jesus Lord at thy birth
	A	Jesus *love* (at) *the* birth
	A-V	Jesus Lord (at) *the* birth
	A-V-I	Jesus Lord (at) thy birth
5 dB	Ref.	But the wine and the song
	A	*If* the *wind in my soul*
	A-V	*Baby* the *wind in* the *sun*
	A-V-I	*Above* the wine and the *soul*
0 dB	Ref.	Like the seasons have all gone
	A	Like the *season's hold* (all) *on*
	A-V	Like the seasons *held* (all) *on*
	A-V-I	Like the seasons *having* (all) gone
-5 dB	Ref.	You gave me love and helped me find the sun
	A	You *give* me love and *help* me *open* (the) *sew*
	A-V	You *give* me love and *help* me *in* the *song*
	A-V-I	You gave me love and *help* me *in* the *song*
-10 dB	Ref.	Sleep in heavenly peace
	A	*Edelweiss edelweiss* (heavenly) (peace)
	A-V	*Sleigh* in *heaven sleigh*
	A-V-I	Sleep in *heaven sleigh*

network to obtain the final output, which has been elaborated in Section 4.3.

B.3 IMU CRNN

In IMU CRNN (see section 4.4 IMU Encoder), the 1D convolutions have {128, 200} channels for each layer, kernel size of 3, and stride of 1. The strides for max-pooling layers are 2. ReLU activation is applied after each layer. Dropout [72] is applied after each convolution and GRU layer with rates of 0.5 and 0.2, respectively. The GRU contains two layers with 60 hidden units in each layer.

B.4 Experiments

For the Voice Activity Detection (VAD) task, we train the model for 50 epochs, evaluate it at the end of each epoch, and select the model with the lowest validation loss. The model is trained using AdamW optimizer [48], with 1×10^{-3} learning rate.

For the Automatic Lyric Transcription (ALT) task, we train the model for 20 epochs using Adam optimizer [43]. Since we adopt the transfer learning paradigm for wav2vec 2.0 and AV-HuBERT, a small learning rate with 1×10^{-5} is used to prevent catastrophic forgetting. For the rest parts of the ALT system, the learning rate ranges from 1×10^{-5} to 5×10^{-4}.

C SOURCE CODE OF OUR PROJECT

The code is attached inside the zip file along with this appendix.

D MORE QUALITATIVE RESULTS

More qualitative results are displayed in Table 9. It is noticed that our MM-ALT system performs better than its audio-only and audio-visual counterparts by reducing word errors.

REFERENCES

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2018. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence* (2018).

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. LRS3-TED: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496* (2018).

[3] Ido Ariav and Israel Cohen. 2019. An end-to-end multimodal voice activity detection using wavenet encoder and residual networks. *IEEE Journal of Selected Topics in Signal Processing* 13, 2 (2019), 265–274.

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).

[5] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems* 33 (2020), 12449–12460.

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[7] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. 2016. End-to-end attention-based large vocabulary speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 4945–4949.

[8] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.

[9] Sakya Basak, Shrutina Agarwal, Sriram Ganapathy, and Naoya Takahashi. 2021. End-to-end lyrics Recognition with Voice to Singing Style Transfer. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 266–270.

[10] Joao Carreira, Skanda Koppula, Daniel Zoran, Adria Recasens, Catalin Ionescu, Olivier Henaff, Evan Shelhamer, Relja Arandjelovic, Matt Botvinick, Oriol Vinyals, et al. 2022. Hierarchical Perceiver. *arXiv preprint arXiv:2202.10890* (2022).

[11] Ccrma.Stanford.Edu. 2018. Smule Sing! 300x30x2 Dataset. Retrieved March 16, 2022 from https://ccrma.stanford.edu/damp/

[12] Shuo-Yiin Chang, Bo Li, Gabor Simko, Tara N Sainath, Anshuman Tripathi, Aäron van den Oord, and Oriol Vinyals. 2018. Temporal modeling using dilated convolution and gating for voice-activity-detection. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5549–5553.

[13] Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. 2021. Multimodal Co-Attention Transformer for Survival Prediction in Gigapixel Whole Slide Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4015–4025.

[14] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[15] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. *Advances in neural information processing systems* 28 (2015).

[16] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622* (2018).

[17] Yu-An Chung and James Glass. 2020. Generative pre-training for speech with autoregressive predictive coding. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3497–3501.

[18] Gerardo Roa Dabike and Jon Barker. 2019. Automatic Lyric Transcription from Karaoke Vocal Tracks: Resources and a Baseline System.. In *Interspeech*. 579–583.

[19] Rebecca Davies, Evan Kidd, and Karen Lander. 2009. Investigating the psycholinguistic correlates of speechreading in preschool age children. *International Journal of Language & Communication Disorders* 44, 2 (2009), 164–174.

[20] Emir Demirel, Sven Ahlbäck, and Simon Dixon. 2020. Automatic lyrics transcription using dilated convolutional neural networks with self-attention. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[21] Emir Demirel, Sven Ahlbäck, and Simon Dixon. 2021. MSTRE-Net: Multistreaming Acoustic Modeling for Automatic Lyrics Transcription. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR*

[*] *2021, Online, November 7-12, 2021*, Jin Ha Lee, Alexander Lerch, Zhiyao Duan, Juhan Nam, Preeti Rao, Peter van Kranenburg, and Ajay Srinivasamurthy (Eds.). 151–158. https://archives.ismir.net/ismir2021/paper/000018.pdf

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[24] Zhiyan Duan, Haotian Fang, Bo Li, Khe Chai Sim, and Ye Wang. 2013. The NUS sung and spoken lyrics corpus: A quantitative comparison of singing and speech. In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 1–9.

[25] Georgi Dzhambazov et al. 2017. *Knowledge-based probabilistic modeling for tracking lyrics in music audio signals*. Ph. D. Dissertation. Universitat Pompeu Fabra.

[26] Hiromasa Fujihara, Masataka Goto, and Jun Ogata. 2008. Hyperlinking Lyrics: A Method for Creating Hyperlinks Between Phrases in Song Lyrics.. In *ISMIR*. 281–286.

[27] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*. 369–376.

[28] Chitralekha Gupta, Emre Yılmaz, and Haizhou Li. 2020. Automatic lyrics alignment and transcription in polyphonic music: Does background music help?. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 496–500.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[30] Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415* (2016).

[31] Takaaki Hori, Shinji Watanabe, and John R Hershey. 2017. Joint CTC/attention decoding for end-to-end speech recognition. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 518–529.

[32] Toru Hosoya, Motoyuki Suzuki, Akinori Ito, Shozo Makino, Lloyd A Smith, David Bainbridge, and Ian H Witten. 2005. Lyrics Recognition from a Singing Voice Based on Finite State Automaton for Music Information Retrieval.. In *ISMIR*. 532–535.

[33] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3451–3460.

[34] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. 2019. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 1440–1444.

[35] Fuyang Huang, Ailing Zeng, Minhao Liu, Qiuxia Lai, and Qiang Xu. 2020. Deepfuse: An imu-aware network for real-time 3d human pose estimation from multiview image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 429–438.

[36] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. 2021. What Makes Multi-modal Learning Better than Single (Provably). *Advances in Neural Information Processing Systems* 34 (2021).

[37] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*. PMLR, 448–456.

[38] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. 2021. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795* (2021).

[39] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*. PMLR, 4651–4664.

[40] Dairoku Kawai, Kazumasa Yamamoto, and Seiichi Nakagawa. 2016. Speech analysis of sung-speech and lyric recognition in monophonic singing. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 271–275.

[41] Fahim Kawsar, Chulhong Min, Akhil Mathur, Alessandro Montanari, Utku Günay Acer, and Marc Van den Broeck. 2018. esense: Open earable platform for human sensing. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. 371–372.

[42] Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. 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, 4835–4839.

[43] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[44] Anna M. Kruspe. 2015. Training Phoneme Models for Singing with "Songified" Speech Data. In *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015, Málaga, Spain, October 26-30, 2015*, Meinard Müller and Frans Wiering (Eds.). 336–342. http://ismir2015.uma.es/articles/34_Paper.pdf

[45] Seungchul Lee, Chulhong Min, Alessandro Montanari, Akhil Mathur, Youngjae Chang, Junehwa Song, and Fahim Kawsar. 2019. Automatic Smile and Frown Recognition with Kinetic Earables. In *Proceedings of the 10th Augmented Human International Conference 2019*. 1–4.

[46] Jinyu Li. 2021. Recent Advances in End-to-End Automatic Speech Recognition. *arXiv preprint arXiv:2111.01690* (2021).

[47] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10012–10022.

[48] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[49] Roya Lotfi, George Tzanetakis, Rasit Eskicioglu, and Pourang Irani. 2020. A comparison between audio and IMU data to detect chewing events based on an earable device. In *Proceedings of the 11th Augmented Human International Conference*. 1–8.

[50] Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2021. End-to-end audio-visual speech recognition with conformers. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7613–7617.

[51] Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264, 5588 (1976), 746–748.

[52] Matt McVicar, Daniel PW Ellis, and Masataka Goto. 2014. Leveraging repetition for improved automatic lyric transcription in popular music. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3117–3121.

[53] Andrew N Meltzoff and M Keith Moore. 1977. Imitation of facial and manual gestures by human neonates. *Science* 198, 4312 (1977), 75–78.

[54] Annamaria Mesaros and Tuomas Virtanen. 2010. Recognition of phonemes and words in singing. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2146–2149.

[55] Gabriel Meseguer-Brocal, Alice Cohen-Hadria, and Geoffroy Peeters. 2019. Dali: A large dataset of synchronized audio, lyrics and notes, automatically created using teacher-student machine learning paradigm. *arXiv preprint arXiv:1906.10606* (2019).

[56] Gabriel Meseguer-Brocal, Alice Cohen-Hadria, and Geoffroy Peeters. 2020. Creating DALI, a large dataset of synchronized audio, lyrics, and notes. *Transactions of the International Society for Music Information Retrieval* 3, 1 (2020).

[57] Chulhong Min, Alessandro Montanari, Akhil Mathur, Seungchul Lee, and Fahim Kawsar. 2018. Cross-modal approach for conversational well-being monitoring with multi-sensory earables. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. 706–709.

[58] Dania Murad, Riwu Wang, Douglas Turnbull, and Ye Wang. 2018. SLIONS: A karaoke application to enhance foreign language learning. In *Proceedings of the 26th ACM international conference on Multimedia*. 1679–1687.

[59] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems* 34 (2021).

[60] Felix Ott, David Rügamer, Lucas Heublein, Tim Hamann, Jens Barth, Bernd Bischl, and Christopher Mutschler. 2022. Benchmarking Online Sequence-to-Sequence and Character-based Handwriting Recognition from IMU-Enhanced Pens. *arXiv preprint arXiv:2202.07036* (2022).

[61] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5206–5210.

[62] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779* (2019).

[63] Stavros Petridis, Themos Stafylakis, Pingehuan Ma, Feipeng Cai, Georgios Tzimiropoulos, and Maja Pantic. 2018. End-to-end audiovisual speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 6548–6552.

[64] Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur. 2018. Semi-orthogonal low-rank matrix factorization for deep neural networks.. In *Interspeech*. 3743–3747.

[65] Sharmin Akther Purabi, Rayhan Rashed, Mirajul Islam, Nahiyan Uddin, Mahmuda Naznin, and ABM Alim Al Islam. 2019. As you are, so shall you move your head: a system-level analysis between head movements and corresponding traits and emotions. In *Proceedings of the 6th International Conference on Networking, Systems and Security*. 3–11.

[66] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al. 2021. SpeechBrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624* (2021).

[67] Siddharth Rupavatharam and Marco Gruteser. 2019. Towards in-ear inertial jaw clenching detection. In *Proceedings of the 1st International Workshop on Earable Computing*. 54–55.

[68] Daniel Seichter, Mona Köhler, Benjamin Lewandowski, Tim Wengefeld, and Horst-Michael Gross. 2021. Efficient rgb-d semantic segmentation for indoor scene analysis. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 13525–13531.

[69] Bidisha Sharma and Ye Wang. 2019. Automatic evaluation of song intelligibility using singing adapted STOI and vocal-specific features. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2019), 319–331.

[70] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184* (2022).

[71] Bowen Shi, Wei-Ning Hsu, and Abdelrahman Mohamed. 2022. Robust Self-Supervised Audio-Visual Speech Recognition. *arXiv preprint arXiv:2201.01763* (2022).

[72] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1 (2014), 1929–1958.

[73] Fei Tao and Carlos Busso. 2020. End-to-end audiovisual speech recognition system with multitask learning. *IEEE Transactions on Multimedia* 23 (2020), 1–11.

[74] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. 2021. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3927–3935.

[75] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. 2017. Total capture: 3d human pose estimation fusing video and inertial sensors. In *Proceedings of 28th British Machine Vision Conference*. University of Surrey, 1–13.

[76] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[77] Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multimodal classification networks hard?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12695–12705.

[78] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi. 2017. Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing* 11, 8 (2017), 1240–1253.

[79] Pengcheng Yang, Boxing Chen, Pei Zhang, and Xu Sun. 2020. Visual agreement regularized training for multi-modal machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 9418–9425.

[80] Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4346–4350.

[81] Chen Zhang, Jiaxing Yu, LuChin Chang, Xu Tan, Jiawei Chen, Tao Qin, and Kejun Zhang. 2021. PDAugment: Data Augmentation by Pitch and Duration Adjustments for Automatic Lyrics Transcription. *arXiv preprint arXiv:2109.07940* (2021).

[82] Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2019. Neural machine translation with universal visual representation. In *International Conference on Learning Representations*.

[83] Pan Zhou, Wenwen Yang, Wei Chen, Yanfeng Wang, and Jia Jia. 2019. Modality attention for end-to-end audio-visual speech recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6565–6569.