# A Benchmark Dataset for Audio-Visual-Electromyography-based Multimodal Speech Recognition

**Abstract**

*This paper introduces a new audio-visual-electromyography (AVE) based multimodal benchmark dataset, featuring a 100-command Mandarin Chinese corpus. Three modalities of speech data, namely audio signals, lip-region images, and six-channel electromyography (EMG) data were collected simultaneously from a diverse group of 100 subjects (29 females and 71 males). Each subject was required to wear the data collection devices and read the whole corpus ten times, with two-second duration for each command in the corpus, contributing to more than 55 hours of multimodal speech data for each modality. Extensive experiments were carried out to investigate the complementary of the different modalities, and provide a baseline fusion paradigm with results indicating the enhancement of the recognition performance for single-modal recognition systems, especially in cross subject and high noise application scenarios. To the best of our knowledge, it is the first phrase-level dataset incorporating three modalities and large-scale Mandarin speech dataset publicly accessible for multimodal researches. We hope it can further promote the innovations in both acoustic and non-acoustic speech recognition fields, providing a solid base for cross-modal learning and making speech interaction a more reliable human machine interaction channel for more users and applications.*

**CCS Concepts**

• ***Computing methodologies*** → ***Speech recognition;*** • ***Human-centered computing*** → *Human computer interaction (HCI);* • ***Information systems*** → ***Database design and models;***

## 1. Introduction

Speech recognition serves as a natural and efficient method for facilitating human-machine interactions. In general, speech recognition can be divided into two categories: automatic speech recognition (ASR) using audio signal [BDMD*07, HDY*12, YD16] and silent speech recognition (SSR) using non-acoustic signals [FEG*08, DSH*10, KKM18]. In recent years, ASR has permeated into daily life, exemplified by the speech-to-text functionality in numerous mobile applications such as Siri by Apple, Alexa by Amazon and WeChat by Tencent. Despite the widespread utility and impressive recognition capability of ASR, its accuracy diminishes in high-noise application environment and some unique situations when acoustic signal cannot be captured.

To overcome these limitations, SSR has experienced accelerated development in sectors such as elderly care, disability assistance, and environments characterized by significant noise. Notably, the advancement of training datasets and deep learning algorithms has spurred substantial progress in specific SSR researches including visual speech recognition (also known as lipreading) [ZZHP14, NYN*14] and surface electromyography (EMG) based speech recognition [MO86, JLA03, WS11]. While on the other hand, visual speech recognition is easily affected by lighting conditions, and the problem of similar mouth shapes for different words cannot be solved well. Surface EMG has the problems that it is easily affected by physiological differences of the users, and the electrode position may shift during speech process. Therefore, utilizing multiple modalities can provide complementary information for speech process, realizing unconstrained speech recognition. Considering a wider range of application scenarios, listing but not limited to the rehabilitation treatment for patients with pronunciation disorders, daily life assistance for the elderly and private communication in the dim or moving indoor scenes and so on, the introduction of audio-visual-electromyography based multimodal speech recognition can overcome the shortcomings of single-model speech recognition systems.

The open-source datasets greatly promote the innovations in speech recognition fields, giving a few examples in the Table 1. Apart from easily accessed audio datasets such as LibriSpeech [PCPK15], SSR datasets also contribute to the algorithm development. For instance, Chuang *et al.* [CZ17] provided a substantial lipreading dataset, comprising a 500-word corpus gathered from television broadcasts. In the later study, Yang *et al.* [YZF*19] presented a naturally-distributed, large-scale Chinese benchmark dataset for lipreading. Some researchers have also proposed multimodal speech recognition datasets. For EMG based speech recognition studies, David Gaddy *et al.* [GK20] collected a 500-utterance dataset of EMG and audio signals from a single speaker. Triantafyllos Afouras *et al.* [ACS*18] introduced and publicly re-

**Table 1:** *Comparison of speech datasets*

| Dataset/Paper | Modality | Language | Corpus Content | Acquisition Source | Number of Speakers | Speaker Indicator |
|---|---|---|---|---|---|---|
| LibriSpeech [PCPK15] | audio | English | about 1000 hours of spoken sentences | LibriVox project with human readers | >1,000 | - |
| THCHS30 [WZ15] | audio | Chinese | 1000 sentences | on-site collection | 40 | - |
| LRW [CZ17] | visual | English | 500 classes of words | television programs | >1,000 | No |
| CAS-VSR-W1k (LRW-1000) [YZF*19] | visual | Chinese | 1,000 classes of words (containing several characters in each) | television programs | >2,000 | No |
| LRS2-BBC [ACS*18] | audio, visual | English | thousands of sentences (containing up to 100 characters in each) | television programs | not given | No |
| EMG-UKA [WJS14] | audio, EMG | English | more than 7,000 utterances | on-site collection | 8 | Yes |
| **Proposed** | **audio, visual, EMG** | **Chinese** | **100 classes of phrases (containing several characters in each)** | **on-site collection** | **100** | **Yes** |

leased a large dataset for audio-visual speech recognition, LRS2-BBC. However, there is no previous work covering three modalities of speech data and involving enough subjects when physiological data is used. Besides, existing large-scale multimodal speech datasets are mainly constructed from television programs [SCSVZ17, YZF*19], lacking of speaker identity, bringing limitations in cross-subject researches and the development of speaker-independent speech recognition system.

To solve the problems of single modality used in speech recognition process, and the lack of large-scale multimodal speech dataset with detailed speaker information and data segmentation, this paper proposes a dataset of audio, visual, and EMG modalities that are directly related to the speech process and provides the benchmark fusion paradigm for multimodal speech recognition. The main contribution of the dataset in this work include:

- The first public dataset in which audio signal, lip image sequence, and facial EMG signal are acquired synchronously;
- The first phrase-level Chinese corpus covering 100 daily life commands, with three to five different Chinese characters in each command;
- The first multimodal speech dataset constructed from 100 Chinese subjects on-site, rather than collected and annotated from public videos.

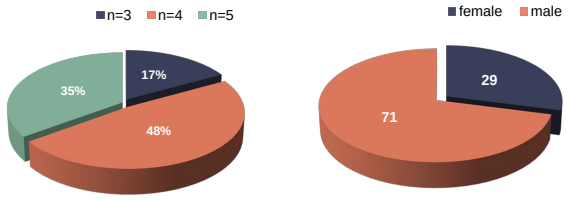
## 2. Dataset Construction

### 2.1. Corpus Design

Based on the five levels of Maslow's Hierarchy of Needs [Mas74], i.e., physiology, safety, belongingness and love, esteem, and self-actualization, a Chinese corpus covering 100 commands was designed to meet the various needs of different users. Considering the spacial needs of speech-impaired individuals and the elderly requiring living assistance, rehabilitation assistance, and daily care, some commands related to medical requirements are also included in the corpus. A few examples of the phrases in the corpus are listed in the Table 2. Each command is a single Mandarin phrase containing 3-5 Chinese words, indicated as independent Mandarin syllables and tones in the phonetic transcription in the Table 2. The distribution of command words is shown in Fig. 1(a). Full corpus content with English translation and phonetic transcription of Mandarin is provided in the supplementary document.

**Table 2:** *Examples of the commands in the corpus*

| Type of needs | Command in English | Phonetic Transcription of Mandarin | Tone |
|---|---|---|---|
| Physiology | I'm Hungry | wo e le | 3 4 5 |
| Safety | Emergency | jin ji hu jiu | 3 2 1 4 |
| Belongingness and Love | I want to have a video chat | wo yao liao shi pin | 3 4 2 4 2 |
| Esteem | I can do it | wo neng xing de | 3 2 2 5 |
| Medical requirements | I keep coughing | wo yi zhi ke sou | 3 4 2 2 4 |
| Medical requirements | My leg aches | wo tui teng | 3 3 2 |
| Medical requirements | How long do I need to stay in hospital | yao zhu yuan duo jiu | 4 4 4 1 3 |
| Medical requirements | My blood pressure is high | wo xue ya gao | 3 3 1 1 |




(a) number of Chinese characters in each demand

(b) number of subjects

**Figure 1:** *Statistic information of the corpus and data collection subject.*

### 2.2. Data Collection Participants

One hundred healthy subjects (29 females and 71 males, aged 20-40 years, mean age: 26.68 years) were recruited in this study, all of which were native speakers of Mandarin Chinese. The data collection experiment was approved by the Research Ethics Committee of Tianjin University. Each subject was provided to read and sign the informed consent form prior to participating in the experiment. The gender distribution of subject group is shown in Fig. 1(b).

### 2.3. Data Acquisition Experimental Procedure

During data collection, subjects were seated in a comfortable chair in a room with normal light and quiet recording environment, assisted with interactive interface showing detailed instructions, as shown in Fig. 2. (a). After the subject wore data collection devices properly, we ran the data acquisition program and started the collection process. A prompt of *read the following command in 2 second* was displayed on the interface, along with command content, guiding the subject to read the phrase correctly in the given period of time. Subjects were asked to refrain from any unnecessary movements during data collection, such as shaking their heads, coughing, yawning, or swallowing. The whole corpus would be read one

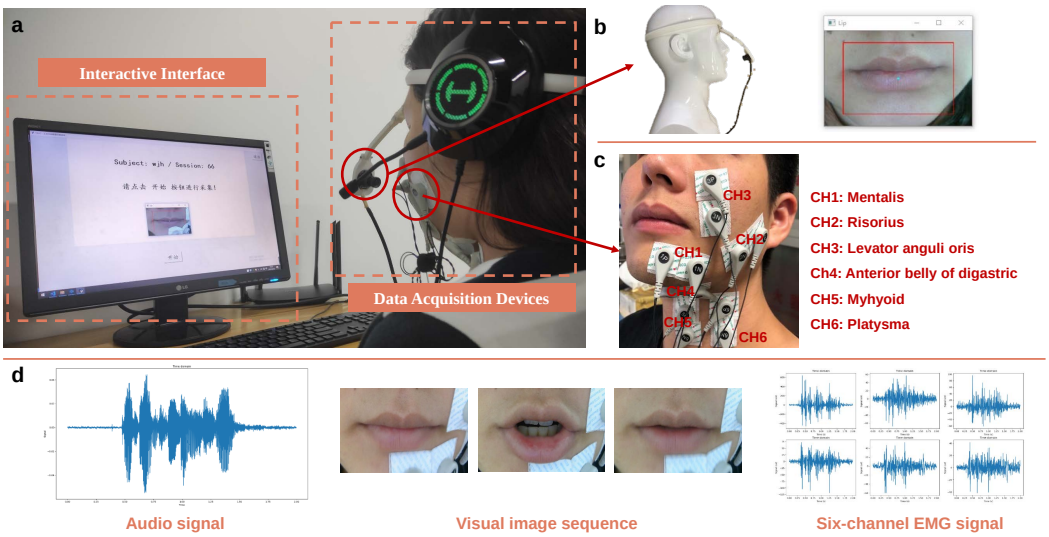



**Figure 2:** *Data acquisition system with both data collection hardware devices and record interface.* ***a*** *The interactive interface used for data collection with subject wearing data collection devices.* ***b*** *The camera used for lip region video capture.* ***c*** *The position of six muscles from which surface electromyography signals were collected.* ***d*** *Data example for each modality.*

time for one round of multimodal speech data collection, while there was 5-second break after data collection of twenty commands. Each subject was required to take ten rounds of data collection, taking about one hour to complete the experiment.

### 2.4. Data Collection Devices

For multimodal speech data acquisition, a head-mounted microphone was used to record audio signals at a sampling rate of $44,100Hz$. RGB camera was used to record lip region videos at a sampling rate of 30 frames per second with resolution of $640 \times 480$. The camera was placed in front of the lip region to record the video data and the distance between the camera and the subject could be adjusted by a 3D-printed fixture, as shown in the Fig. 2.(b). Surface EMG data was collected from six muscles by NSW308M bipolar EMG system (Neuracle Technology Co., Ltd) to record six-channel EMG data at a sampling rate of $1,000Hz$. Six pairs of electrodes were attached on the surface of six facial and neck muscles for speech activity, as shown in Fig. 2.(c). The reference electrode was placed upon the collarbone to record the voltage brought by the human body as a baseline. Electrode impedance was kept below $10k\Omega$ during recording.

### 2.5. Dataset Archive Information

Information for all subjects is listed in a *Sub_info.txt* file. For each subject, there are three factors including *Subject Index*, *Gender*, and *Age*. The data file is stored into four levels. The first level contains three folders including *audio*, *video*, and *EMG* indicating each modality. Each folder includes 100 different subject folders, which is the second level storage. The third-level subject folder keeps 10 sessions of data in 10 different folders, and in the last level, each session folder has 101 pieces of single-modal speech data corresponding to each command and one blank command data. The file name is used as data label in the following experiments. Information for all labels is listed in a *label_info.txt* file. The dataset contains more than 1000 utterances from each speaker and a total of 99500 valid samples for one modality of speech data, counting for nearly 55.3 hours.

### 2.6. Data Pre-processing

Simple steps of data pre-processing can be performed to improve the data quality for the following researches. We do not recommend any data pre-processing for audio signal because of the high-quality data acquisition of the head-mounted microphone. Framing is performed to the acquired lip region videos and the image sequences are further converted to gray-scale, followed by cropping a region of interest (ROI) to exclude unimportant facial parts and resizing the region to a set size of $88 \times 88$.

Interference such as peak amplitude, DC offset, and $50Hz$ power frequency noise persists in the EMG data, and filtering is advised to improve the data quality. To filter out the $50Hz$ power frequency noise as well as its odd harmonics at $150Hz$, $250Hz$, and $350Hz$, we employ a second-order Butterworth notch filter. A $10-400Hz$ Butterworth bandpass filter is then applied in light of the discovery that the most useful component of EMG data lies in the frequency range between $10Hz$ and $400Hz$ [LXZ14]. A valid example set of the multimodal speech data for one command in the corpus, including simultaneously collected audio data, pre-processed six-channel EMG data, and lip region images framed from videos centered at the mouth from one subject is shown in the Fig. 2. (d).

### 2.7. Basic Feature Extraction

Mel spectral features are made to glean information about the speech from the standpoint of human auditory characteristics. Fol-

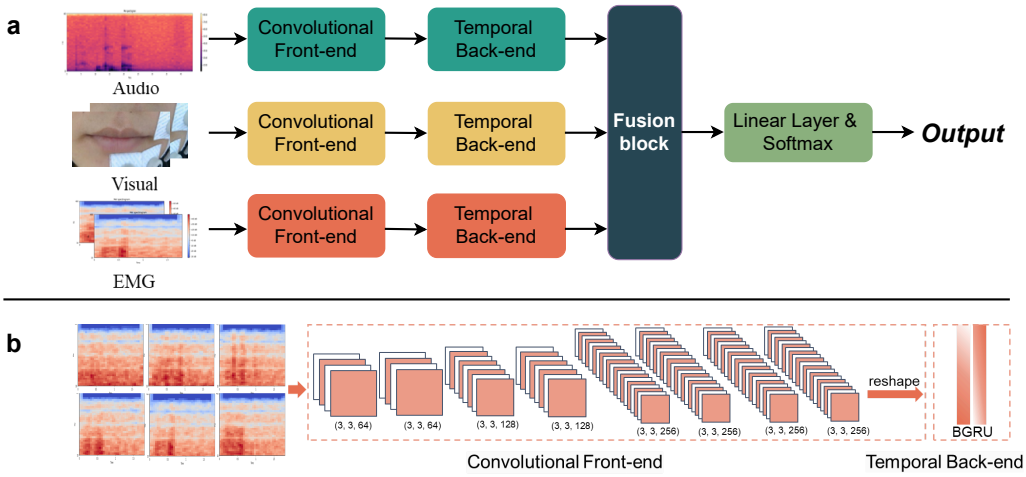



**Figure 3:** *Fusion network architecture.* ***a*** *The fusion pipeline with three modalities of speech data.* ***b*** *Detailed architecture for EMG branch.*

lowing the successful implementation of Mel Frequency Cepstral Coefficient (MFCC) in automatic speech recognition systems, log Mel-frequency spectral coefficients (MFSC) is proposed without the final discrete cosine transform (DCT) in the end of the MFCC feature extraction, preserving a higher feature dimension [Moh14]. Therefore, we extract the $60 \times 64$ MFSC features for each piece of raw audio sample. Similar to audio data, MFSC features with dimension of $6 \times 36 \times 36$ are extracted as basic feature for the filtered EMG data.

## 3. Preliminary Fusion Experiments

We adopt the commonly used methods for single-modal speech recognition and conventional fusion paradigm in this section and carry out speech recognition experiments based on proposed dataset. A brief overview of the multimodal speech recognition network used in this work is shown in Fig. 3. (a).

### 3.1. Audio-Visual-EMG Based Speech Recognition Architecture

**Single-stream speech recognition model:** For audio and video feature streams, the front-end module developed in previous work [PSM*18, MPP21, LXY21] has been adopted. We initially employ the 2D and 3D Convolutional layer as the front-end encoder for audio and lip region image sequence, respectively, attached by the ResNet 18 model. A temporal back-end has been adopted here for temporal information extraction. The 2-layer Bidirectional Gated Recurrent Unit (BGRU) model and Transformer encoder have been selected, considering the wide and successful implementation of these two algorithms in previous work. The hidden size is 512 for BGRU layers and the default dimension of 2048 is used for the feed forward network in Transformer encoder.

With the findings of previous work on EMG-based silent speech recognition [WZZ*21, ZCW*23], we utilize deep CNN architecture as multi-channel spatial feature extractor in this study, as demonstrated in the Fig. 3. (b). In practice, 2D Convolutional layers with kernel size of $3 \times 3$ and channel numbers of $(64, 64, 128, 128, 256)$ are utilized for front-end feature extraction. The same temporal back-end selection is used in EMG stream as well. The linear layer and softmax function are used for phrase classification for single stream speech recognition.

**Multimodal fusion architecture:** A basic audio-visual-EMG based multimodal speech recognition network is proposed, based on single-stream models designed above. As one of the most popular fusion paradigm, feature fusion is adopted in this network. It has demonstrated exceptional performance in audiovisual speech recognition tasks [MPP21, HK13], where multiple hidden layers make it simple to combine temporal and spatial information from different modalities. Single-stream model serves as spatio-temporal feature extractor of each modality, keeping the unique speech information and passing it to the concatenation layer of the fusion block. After feature concatenation, the fused multimodal speech feature is further modelled by the temporal neural network for cross-modal temporal feature extraction, which is the same as the temporal back-end used in each single stream. To be noticed, the temporal back-end used in the single stream during feature fusion is BGRU network, because of the superior recognition performance in single stream recognition experiments.

### 3.2. Training Strategies

For efficient model training and convergence, we conducted pre-training on the three single-modal speech recognition models. The implementation was based on PyTorch and the models were trained on the server with single NVIDIA Titan V GPU of 24 GB memory. Adam optimizer was used without weight decay. The initial learning rate was 0.0003 for BGRU back-end. A larger learning rate of 0.001 was used for Transformer encoder, with a learning rate decay after twenty to thirty epochs, according to different modalities. We used cross entropy loss as a loss function in the training process to measure the performance of the model. After single-modal pre-

training, model parameters were frozen and loaded in the fusion model training.

### 3.3. Experimental Settings

**Cross-subject speech recognition:** The whole dataset was used in the experiment. To be specific, the data collected from the first 70 subjects were used as training set, the data collected from the following 10 subjects were used as validation set and data collected from the last 20 subjects were put into test set. Since there was no overlap of subjects, we evaluated the cross-subject recognition ability of the multimodal fusion network in the following experiments, providing the recognition accuracy of each modality and fusion model using validation set and test set.

**High-noise speech recognition:** As demonstrated above, the audio signal is of high quality. To verify the improvement of multimodal fusion on automatic speech recognition, different intensities of Gaussian noise were added to the audio signal in the test phase directly for recognition. Signal-to-Noise Ratio (SNR) was used to measure the noise level that added to the test audio.

### 3.4. Recognition results

*a) Single-modal based speech recognition experiments*

The recognition results of single-modal speech recognition model are shown in Table 3, under different audio SNRs in the test phase. For audio and EMG based speech recognition models, the BGRU back-end shows better recognition performance, while it is the opposite for lipreading model. Besides, signal differences caused by physiological conditions of different subjects lead to the relatively low recognition performance in this cross-subject experiment using EMG speech recognition model. Since EMG signal can be seldom affected by the surrounding environment, it can be a good compensation for direct speech information when integrated with other modalities. With the decrease in the SNR, the recognition accuracy drops greatly using automatic speech recognition method. It gets worse when there are unseen speakers in the validation and test sets, making it more difficult to maintain a stable recognition result. When SNR falls below $0dB$, both EMG and visual modalities can provide sufficient speech information for recognition, which is the initial consideration for the use of multimodal fusion paradigm to improve the speech recognition performance in this work.

*b) Multimodal speech recognition experiments*

Experiments of multimodal fusion network under different test audio SNRs are carried out, as shown in Table 4.

The model integrated with three modalities has significantly improved the recognition accuracy under different noise levels, compared with the audio model. Especially when using the Transformer encoder for cross-modal temporal feature extraction, the recognition accuracy of the fusion model degrades more slowly, reaching a recognition accuracy above 90% When the test audio SNR drops to $-10dB$. At the same time, the recognition differences for unseen speakers in validation and test sets are much smaller, compared to single streams, indicating robust recognition ability for unseen speakers.

**Table 3:** *Recognition results of single-modal streams*

| Modality | Backend | SNR of Test Audio Data | Validation Set | Test Set |
|---|---|---|---|---|
| Audio | BGRU | N/A | 99.47 | 99.45 |
| | | 10 dB | 97.53 | 98.70 |
| | | 5 dB | 91.65 | 95.40 |
| | | 0 dB | 70.82 | 80.45 |
| | | -5 dB | 36.41 | 45.47 |
| | | -10 dB | 6.32 | 8.78 |
| | Transformer Encoder | N/A | 99.30 | 99.36 |
| | | 10 dB | 96.92 | 98.49 |
| | | 5 dB | 86.98 | 93.98 |
| | | 0 dB | 59.56 | 72.16 |
| | | -5 dB | 28.61 | 35.33 |
| | | -10 dB | 6.31 | 8.13 |
| EMG | BGRU | N/A | 77.53 | 75.53 |
| | Transformer Encoder | N/A | 61.34 | 63.54 |
| Visual | BGRU | N/A | 92.45 | 95.09 |
| | Transformer Encoder | N/A | 98.71 | 98.55 |

**Table 4:** *Recognition results of fusion network*

| Modality | Backend | SNR of Test Audio Data | Validation Set | Test Set |
|---|---|---|---|---|
| Fusion | BGRU | clean | 98.82 | 99.18 |
| | | 10 dB | 98.46 | 99.01 |
| | | 5 dB | 97.74 | 98.54 |
| | | 0 dB | 95.93 | 97.17 |
| | | -5 dB | 92.52 | 94.36 |
| | | -10 dB | 86.41 | 88.57 |
| | Transformer Encoder | clean | 99.70 | 99.87 |
| | | 10 dB | 99.58 | 99.81 |
| | | 5 dB | 99.12 | 99.64 |
| | | 0 dB | 98.15 | 99.12 |
| | | -5 dB | 96.12 | 97.81 |
| | | -10 dB | 93.00 | 95.51 |

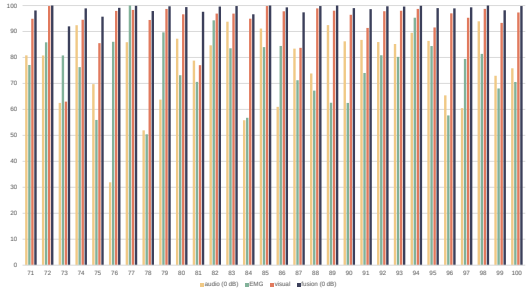
*c) Recognition performance for single subject using different modalities and fusion network*

The recognition results for each subject in the validation and test datasets are presented in the Fig. 4 when test audio SNR drops to $0dB$, $-5dB$ and $-10dB$, respectively.

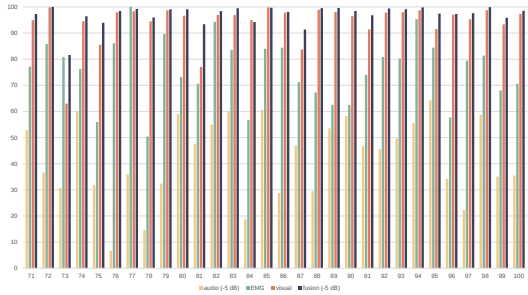
There is obvious physiological and appearance variation across subjects, leading to the fluctuations of recognition accuracy. Especially when using single stream data, these is more distinguished differences of recognition performance, compared to fusion network. Experimental results confirm that the integration of multimodal speech data not only improves the recognition results under severe audio noise, but also smooths the differences brought by subjects, especially when test audio SNR drops to $-10dB$, as shown in the Fig 4. (c). It is particularly important for practical applications where new users can obtain stable interaction experience without calibration of recognition model.

### 3.5. Further Discussion

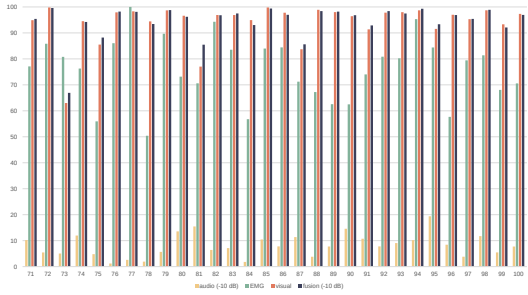
The multilevel and rich-grained semantic information brought by audio, visual and EMG data compensates for the limitations of single-modal speech recognition system in cross subject and high noise application scenarios. Nevertheless, it is still noticed that conventional fusion paradigm may have worse recognition than single-



(a) Test audio SNR 0dB



(b) Test audio SNR -5dB



(c) Test audio SNR -10dB

**Figure 4:** *Recognition results for unseen subjects using different modalities and the fusion network.*

modal streams when one of the modalities is severely corrupted. For example, when SNR of test audio data falls to $-10dB$, the fusion network obtains relatively lower recognition accuracy of 93% and 95.51% using Transformer back-end for validation and test set, which are lower than results brought by visual speech recognition network, i.e., 98.71% and 98.55% for validation and test sets using the same temporal back-end. From this perspective, advanced fusion method shall be designed to make the best use of multiple modalities of speech information during recognition process and resist the noise interference in single modalities.

### 3.6. Potential research fields

Proposed multimodal speech dataset contains high-quality and aligned speech data collected from diverse speakers, providing a vital database for a wide range of researches and applications. Several possible research topics are given below, and we believe more inspiring work can be carried out based on it.

**Multimodal fusion strategy.** Conventional fusion strategies like feature fusion [MPP21, SSL22], decision fusion [YZK21] and the combination of them [LLY21] have been widely explored in multimodal speech recognition researches. Furthermore, different learning methods such as self-supervised learning [SHLM22, SHM22] have been used to improve the representation ability of single modalities and modality-agnostic linguistic features. Proposed AVE multimodal speech dataset involves three heterogeneous speech modalities, bringing great chance in the discovery of single-modal speech information representation and cross-modal correlations. The different feature dimension and sparsity of semantic information lying in one-dimensional signal (i.e., audio and EMG) and high-dimensional image sequence (visual data) can be explored and mitigated during fusion process. With the considerable amount of multimodal data, both supervised and self-supervised learning approaches can be developed, as well as multitask learning and cross-model knowledge distillation models and so on.

**Single-modal speech enhancement.** Multimodal speech enhancement emphasizes research focus on improving the speech quality and perception [MTZ*21,RXW*21]. Recent studies mainly utilize audio and visual speech information in the multimodal speech enhancement, using disentanglement learning [CRG21],generative approach [RFG23], as well as unsupervised learning [SAP21]. With the proposed AVE multimodal speech dataset, we can further investigate the potential of speech enhancement using audio and EMG data. Since EMG signal reflects the muscle move-ment during speech directly, it describes the connection between human articulation movement and speech content. The similar sig-nal waveform of audio and EMG data also lays good foundation for mutual information extraction and speech enhancement when au-dio noise or speaker differences occur. Furthermore, both audio and EMG data can be used to improve the lipreading ability when there is large noise and the lighting condition is undesirable. With the knowledge distilled from the other two modalities, single-modal speech enhancement can obtain more satisfactory performance and we can further use just one modality in real-world applications.

## 4. Conclusion

In this paper, we proposed a novel audio-visual-EMG based multimodal Chinese phrase-level benchmark dataset, named AVE, for speech recognition researches. Specifically, we designed a Chinese corpus covering 100 classes of short commands and collected data from a total of 100 subjects. With this new dataset, we compared the performance of models fusing three modalities with single-modal speech recognition models under different test audio SNRs. Experimental results confirm that the fusion network can effectively improve the recognition accuracy and robustness for unseen speakers, and there is large improvement space for advanced fusion approaches with better recognition performance in complex application environment. Therefore, proposed AVE multimodal speech dataset is a promising platform for further researches and innovations in speech recognition and multimodal learning, and we hope

it can be a useful tool to explore the mechanism of human speech perception and interaction.

## References


[ACS*18] AFOURAS T., CHUNG J. S., SENIOR A., VINYALS O., ZISSERMAN A.: Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence 44*, 12 (2018), 8717–8727. 1, 2

[BDMD*07] BENZEGHIBA M., DE MORI R., DEROO O., DUPONT S., ERBES T., JOUVET D., FISSORE L., LAFACE P., MERTINS A., RIS C., ET AL.: Automatic speech recognition and speech variability: A review. *Speech communication 49*, 10-11 (2007), 763–786. 1

[CRG21] CARBAJAL G., RICHTER J., GERKMANN T.: Disentanglement learning for variational autoencoders applied to audio-visual speech enhancement. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (2021), IEEE, pp. 126–130. 6

[CZ17] CHUNG J. S., ZISSERMAN A.: Lip reading in the wild. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13* (2017), Springer, pp. 87–103. 1, 2

[DSH*10] DENBY B., SCHULTZ T., HONDA K., HUEBER T., GILBERT J. M., BRUMBERG J. S.: Silent speech interfaces. *Speech Communication 52*, 4 (2010), 270–287. 1

[FEG*08] FAGAN M. J., ELL S. R., GILBERT J. M., SARRAZIN E., CHAPMAN P. M.: Development of a (silent) speech recognition system for patients following laryngectomy. *Medical engineering & physics 30*, 4 (2008), 419–425. 1

[GK20] GADDY D., KLEIN D.: Digital voicing of silent speech. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2020), pp. 5521–5530. 1

[HDY*12] HINTON G., DENG L., YU D., DAHL G. E., MOHAMED A.-R., JAITLY N., SENIOR A., VANHOUCKE V., NGUYEN P., SAINATH T. N., ET AL.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine 29*, 6 (2012), 82–97. 1

[HK13] HUANG J., KINGSBURY B.: Audio-visual deep learning for noise robust speech recognition. In *2013 IEEE international conference on acoustics, speech and signal processing* (2013), IEEE, pp. 7596–7599. 4

[JLA03] JORGENSEN C., LEE D. D., AGABONT S.: Sub auditory speech recognition based on emg signals. In *Proceedings of the International Joint Conference on Neural Networks, 2003.* (2003), vol. 4, IEEE, pp. 3128–3133. 1

[KKM18] KAPUR A., KAPUR S., MAES P.: Alterego: A personalized wearable silent speech interface. In *23rd International conference on intelligent user interfaces* (2018), pp. 43–53. 1

[LLY21] LIU H., LI W., YANG B.: Robust audio-visual speech recognition based on hybrid fusion. In *2020 25th International Conference on Pattern Recognition (ICPR)* (2021), IEEE, pp. 7580–7586. 6

[LXY21] LIU H., XU W., YANG B.: Audio-visual speech recognition using a two-step feature fusion strategy. In *2020 25th International Conference on Pattern Recognition (ICPR)* (2021), IEEE, pp. 1896–1903. 4

[LXZ14] LYU M., XIONG C., ZHANG Q.: Electromyography (emg)-based chinese voice command recognition. In *2014 IEEE International Conference on Information and Automation (ICIA)* (2014), IEEE, pp. 926–931. 3

[Mas74] MASLOW A.: *A theory of human motivation*. Lulu. com, 1974. 2

[MO86] MORSE M. S., O'BRIEN E. M.: Research summary of a scheme to ascertain the availability of speech information in the myoelectric signals of neck and head muscles using surface electrodes. *Computers in biology and medicine 16*, 6 (1986), 399–410. 1

[Moh14] MOHAMED A.-R.: *Deep Neural Network Acoustic Models for ASR*. PhD thesis, University of Toronto Toronto, Canada, 2014. 4

[MPP21] MA P., PETRIDIS S., PANTIC M.: End-to-end audio-visual speech recognition with conformers. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2021), IEEE, pp. 7613–7617. 4, 6

[MTZ*21] MICHELSANTI D., TAN Z.-H., ZHANG S.-X., XU Y., YU M., YU D., JENSEN J.: An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 29* (2021), 1368–1396. 6

[NYN*14] NODA K., YAMAGUCHI Y., NAKADAI K., OKUNO H. G., OGATA T.: Lipreading using convolutional neural network. In *fifteenth annual conference of the international speech communication association* (2014). 1

[PCPK15] PANAYOTOV V., CHEN G., POVEY D., KHUDANPUR S.: Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (2015), IEEE, pp. 5206–5210. 1, 2

[PSM*18] PETRIDIS S., STAFYLAKIS T., MA P., CAI F., TZIMIROPOULOS G., PANTIC M.: End-to-end audiovisual speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (2018), IEEE, pp. 6548–6552. 4

[RFG23] RICHTER J., FRINTROP S., GERKMANN T.: Audio-visual speech enhancement with score-based generative models. In *Speech Communication; 15th ITG Conference* (2023), pp. 275–279. doi:10.30420/456164054. 6

[RXW*21] RAMESH K., XING C., WANG W., WANG D., CHEN X.: Vset: A multimodal transformer for visual speech enhancement. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2021), IEEE, pp. 6658–6662. 6

[SAP21] SADEGHI M., ALAMEDA-PINEDA X.: Mixture of inference networks for vae-based audio-visual speech enhancement. *IEEE Transactions on Signal Processing 69* (2021), 1899–1909. doi:10.1109/TSP.2021.3066038. 6

[SCSVZ17] SON CHUNG J., SENIOR A., VINYALS O., ZISSERMAN A.: Lip reading sentences in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 6447–6456. 2

[SHLM22] SHI B., HSU W.-N., LAKHOTIA K., MOHAMED A.: Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184* (2022). 6

[SHM22] SHI B., HSU W.-N., MOHAMED A.: Robust self-supervised audio-visual speech recognition. *arXiv preprint arXiv:2201.01763* (2022). 6

[SSL22] SONG Q., SUN B., LI S.: Multimodal sparse transformer network for audio-visual speech recognition. *IEEE Transactions on Neural Networks and Learning Systems* (2022). 6

[WJS14] WAND M., JANKE M., SCHULTZ T.: The emg-uka corpus for electromyographic speech processing. In *INTERSPEECH* (2014), pp. 1593–1597. 2

[WS11] WAND M., SCHULTZ T.: Session-independent emg-based speech recognition. In *Biosignals* (2011), pp. 295–300. 1

[WZ15] WANG D., ZHANG X.: Thchs-30: A free chinese speech corpus. *arXiv preprint arXiv:1512.01882* (2015). 2

[WZZ*21] WU J., ZHAO T., ZHANG Y., XIE L., YAN Y., YIN E.: Parallel-inception cnn approach for facial semg based silent speech recognition. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (2021), IEEE, pp. 554–557. 4

[YD16] YU D., DENG L.: *Automatic speech recognition*, vol. 1. Springer, 2016. 1

[YZF*19] YANG S., ZHANG Y., FENG D., YANG M., WANG C., XIAO J., LONG K., SHAN S., CHEN X.: Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. In *2019 14th IEEE*

*international conference on automatic face & gesture recognition (FG 2019)* (2019), IEEE, pp. 1–8. 1, 2

[YZK21] YU W., ZEILER S., KOLOSSA D.: Fusing information streams in end-to-end audio-visual speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2021), IEEE, pp. 3430–3434. 6

[ZCW*23] ZHANG Y., CAI H., WU J., XIE L., XU M., MING D., YAN Y., YIN E.: Emg-based cross-subject silent speech recognition using conditional domain adversarial network. *IEEE Transactions on Cognitive and Developmental Systems* (2023). 4

[ZZHP14] ZHOU Z., ZHAO G., HONG X., PIETIKÄINEN M.: A review of recent advances in visual speech decoding. *Image and vision computing 32*, 9 (2014), 590–605. 1