ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10890482

M-BEST-RQ: A Multi-Channel Speech Foundation Model for Smart Glasses

Yufeng Yang[1*,2], Desh Raj[1], Ju Lin[1], Niko Moritz[1], Junteng Jia[1], Gil Keren[1],
Egor Lakomkin[1], Yiteng Huang[1], Jacob Donley[1], Jay Mahadeokar[1], Ozlem Kalinli[1]
[1]Meta, USA [2]The Ohio State University, USA
yang.5662@osu.edu, desh@meta.com

***Abstract*—The growing popularity of multi-channel wearable devices, such as smart glasses, has led to a surge of applications such as targeted speech recognition and enhanced hearing. However, current approaches to solve these tasks use independently trained models, which may not benefit from large amounts of unlabeled data. In this paper, we propose M-BEST-RQ, the first multi-channel speech foundation model for smart glasses, which is designed to leverage large-scale self-supervised learning (SSL) in an array-geometry agnostic approach. While prior work on multi-channel speech SSL only evaluated on simulated settings, we curate a suite of real downstream tasks to evaluate our model, namely (i) conversational automatic speech recognition (ASR), (ii) spherical active source localization, and (iii) glasses wearer voice activity detection, which are sourced from the MMCSG and EasyCom datasets. We show that a general-purpose M-BEST-RQ encoder is able to match or surpass supervised models across all tasks. For the conversational ASR task in particular, using only 8 hours of labeled speech, our model outperforms a supervised ASR baseline that is trained on 2000 hours of labeled data, which demonstrates the effectiveness of our approach.**

***Index Terms*—Beamforming, BEST-RQ, multi-channel, self-supervised learning, smart glasses.**

I. Introduction

With the growing popularity and adoption of multi-channel smart wearable devices such as smart glasses, there are several new use cases related to the spatial understanding of audio and speech. These include automatic speech recognition (ASR) for smart assistants, enhanced hearing, etc [1]. Smart wearable devices usually consist of multi-channel audio input and involve the wearer's interaction with the device and surrounding objects or participants. However, because of the limited annotated data from such devices, the use of multi-channel inputs is often limited to traditional signal processing. Furthermore, different use cases are usually addressed using separate models that do not make use of the knowledge from other tasks. Self-supervised learning (SSL) has been shown to be effective on low-resource tasks with representations learned from unlabeled data [2]–[7], and "foundation models" trained using such methods have recently outperformed supervised models [8]. Framing the above problem in the low-resource context, our objective in this work is to build the first foundation model specifically for tasks based around wearable devices such as smart glasses.

Most existing work on speech SSL has focused on single-channel inputs [3]–[6]. For multi-channel SSL, while researchers have proposed methods such as Spatial HuBERT [9], multi-channel AV-wav2vec2 [10], and UniX-Encoder [11], these models have only been evaluated in limited settings of simulated data or fixed array-geometry. Different from these, we want to build a foundation model that be fine-tuned on several downstream tasks and can work across wearable devices with different numbers of microphones and array geometries. Our key insight to achieve device-agnosticity is to use multiple super-directivity beamformers to convert "channels" to a fixed number of "directions" which can be processed by the neural encoder [12]. For the SSL backend, we propose a multi-channel extension of BEST-RQ [13], since it is conceptually simple and has been shown to outperform other methods such as wav2vec 2.0 [14]. We refer to the resulting model as **M-BEST-RQ**.

Following the conventional paradigms, we first pre-train the M-BEST-RQ encoder using masked estimation methods on a combination of large-scale synthetic data simulated from public datasets (such as LibriSpeech [15] and Libri-Light [16]) and real multi-channel data from the Project Aria glasses [1]. We then use the encoder for supervised fine-tuning and evaluation on several downstream tasks: conversational ASR (C-ASR), spherical active source localization (S-ASL), and glasses wearer voice activity detection (W-VAD). On the C-ASR task (formulated using the recently published MMCSG dataset [17]), M-BEST-RQ achieves 20.1%/28.1% word error rate (WER) for self/other speaker using only 8 hours of labeled speech, outperforming an ASR baseline trained on 2k hours labeled data. We curate the S-ASL and W-VAD tasks using the EasyCom dataset [18], which is recorded on a different device than Aria. On these tasks, our audio-only M-BEST-RQ model matches or outperforms baselines trained with audio-visual modalities, indicating that M-BEST-RQ is a generic foundation model that can work for several downstream tasks on different devices.

II. M-BEST-RQ

Given raw speech $\mathbf{X} \in \mathbb{R}^{T\times M}$ containing T samples collected on M microphones, the problem is to learn a function f (the foundation model) which converts $\mathbf{X}$ into task-and-array-agnostic high-dimensional representations $\mathbf{F}^{T'\times D}$, where T' is usually down-sampled from T. At a high-level, there are two questions that must be answered in order to achieve this: (i) How do we make the foundation model f invariant to the number of microphone channels M and their geometry? (ii) How do we learn f such that it is "generic," i.e., it can perform well on any downstream task $\mathcal{T}$?

A. Array Invariance with Fixed Beamformers

For general-purpose multi-channel devices, array geometry invariance may be achieved using neural methods such as cross-channel attention [19] at the cost of increased computation. However, this solution disregards the fact that all the array geometries, despite being different, are situated on wearable devices and are used for wearer-related tasks. With this assumption in place, we can instead make use of device-specific beamformers to convert an arbitrary number of input "channels" M to a fixed number of input "directions" K (which equals to 12), as shown in Fig. 1. Formally, we have

$$\mathbf{F} = f(\mathbf{X}) = (f_{\text{enc}} \circ f_{\text{bf}})(\mathbf{X}) = f_{\text{enc}}\left(f_{\text{bf}}(\mathbf{X})\right), \quad (1)$$

where $f_{\text{bf}}(\mathbf{X}) = \mathbf{X}' \in \mathbb{R}^{T\times K}$ is fixed, and f_{enc} is parameterized by a neural network. Here, $\mathbf{X}'$ is invariant to the array geometry, and represents a collection of signals from K directions.

For our function f_{bf}, we use non-linearly constrained minimum-variance (NLCMV) beamforming [12], [20]. Given the mouth-

*Work done during internship at Meta

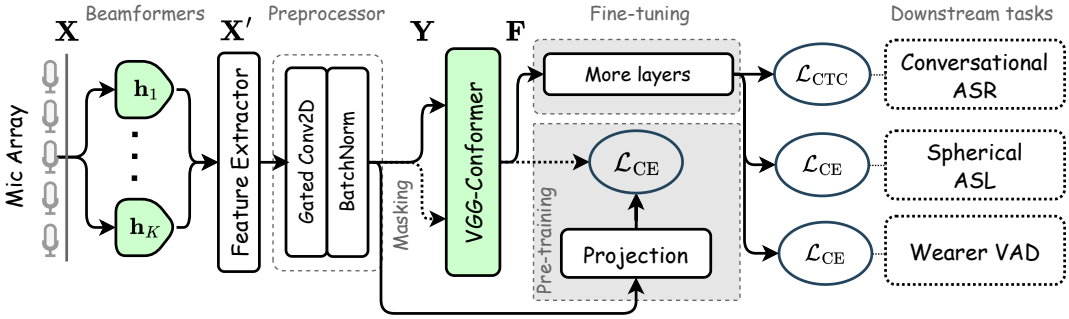


Fig. 1: System architecture of M-BEST-RQ and downstream tasks.

directed and far-field acoustic transfer functions (ATFs), the beamforming weights $\mathbf{h}_k(j\omega)$ of each steering direction $k \in \{1, \ldots, K\}$ are obtained by minimizing

$$\mathbf{h}^H(j\omega)\left[\mathbf{\Phi}_{dd}(j\omega) + \underbrace{\phi_{pp}(\omega)\sum_{n=1}^{N}\alpha_{p,n}\cdot\mathbf{g}_n(j\omega)\mathbf{g}_n^H(j\omega)}_{\text{soft control of null directions.}}\right]\mathbf{h}(j\omega), \tag{2}$$

which is subject to the linear equality and nonlinear inequality constraints, and they are simplified to

$$\begin{cases}\mathbf{h}^H(j\omega)\mathbf{g}(j\omega) = 1,\\ c(\omega) \triangleq \underbrace{\mathbf{h}^H(j\omega)\mathbf{\Psi}(j\omega)\mathbf{h}(j\omega) \le 0}_{\text{constraint on white noise gain.}},\end{cases} \tag{3}$$

where $\mathbf{\Phi}_{dd}(j\omega)$ is the covariance matrix of diffuse noise, and

$$\mathbf{\Psi}(j\omega) \triangleq \mathbf{I} - \mathbf{g}_n(j\omega)\mathbf{g}_n^H(j\omega)\cdot M \Big/ \left[\sum_{m=1}^{M}|G_m(j\omega)|^2\right]. \tag{4}$$

The $G_m(j\omega)$ are measured channel responses from the target speech source to the m-th microphone (ATFs). N is the number of point noise sources, $\phi_{pp}(\omega)$ is the power spectral density of point noise, $\alpha_{p,n}$ is the n-th point noise weight, $\mathbf{g}_n(j\omega)$ is the channel response from the n-th point noise source, and $\mathbf{I}$ is the identity matrix. The nature of the directive signal enables the model to learn array-geometry agnostic representations, and with ATFs available, the NLCMV beamforming can be applied to any new device, i.e., $\mathbf{X}' = \mathbf{h}_k(j\omega)\mathbf{X}$.

B. Task Invariance with BEST-RQ

We extend the BEST-RQ [13] to work with multi-channel inputs by affixing a multi-channel projection module in front. Given $\mathbf{X}'$, we extract the corresponding log-Mel filterbank features, and then project the K channels into a single channel using a gated convolution followed by batch-normalization, thus resulting in latent representations $\mathbf{Y}$ which contain information from all K directions. These representations are provided as input to a VGG-Conformer encoder (f_{enc}), which outputs semantic embeddings $\mathbf{F}$.

Similar to [13], we train the encoder by masking random chunks of $\mathbf{Y}$ before feeding to f_{enc}. An unmasked $\mathbf{Y}$ is projected using a randomly initialized quantizer into discrete labels, and the pre-training objective is to predict the labels corresponding to the masked regions using cross-entropy loss, $\mathcal{L}_{\text{CE}}$. During the fine-tuning stage, we add additional layers to the output of the VGG-Conformer encoder and train it on labeled data for different downstream tasks with different loss functions, as shown in Fig. 1. Our conjecture is that the masking $\mathbf{Y}$ in the pre-training stage imparts semantic as well as directional understanding of multi-channel speech to the M-BEST-RQ encoder, which would enable it to work well on several tasks.

III. Experimental Setup

A. Datasets

We simulated 7-channel LibriSpeech (LS) and Libri-Light (LL) datasets according to the array configuration of the Project Aria glasses, based on the original LibriSpeech [15] and Libri-Light [16] datasets. For this, we first segmented the long utterances into shorter segments (between 0.5 and 10 seconds) based on forced alignments[1]. We then generated 100k room impulse responses (RIRs) for the Aria microphone array, and used these to simulate multi-channel, multi-speaker conversations between a wearer, a participant, and a distractor speaker. The simulation process is the same as the "train-from-scratch" baseline of the MMCSG challenge [17]. The duration of the simulated multi-channel LS and LL datasets are about 2k and 140k hours, respectively, with each utterance being 12±5 seconds.

In addition, we also used ∼800 hours of real, in-house, multi-channel data collected using the Aria glasses to investigate the impact of real data in pre-training. We downsampled these recordings from 48 kHz into 16 kHz and randomly segmented the recording into 12±5 second segments, resulting in segments in the range [2, 30] seconds. We refer to this dataset as RD.

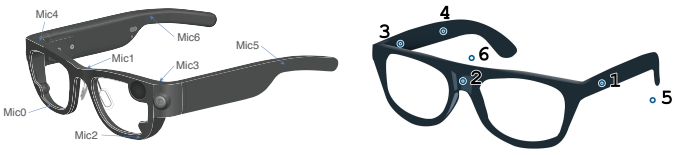
For fine-tuning and evaluation, we curated 3 downstream tasks based on the MMCSG [17] and EasyCom [18] datasets. MMCSG is released as part of the CHiME-8 challenge[2] focusing on transcribing natural conversations between two speakers, recorded on Aria glasses. The duration of training, development, and evaluation data is 8.5, 8.4, and 9.4 hours, respectively. EasyCom is a dataset of multi-talker conversations in noisy environments with egocentric video recorded on a pair of augmented-reality (AR) glasses with a different number and array of microphones from the Aria glasses. The duration of the dataset is about 5.3 hours. Details of the microphone positions of two glasses are shown in Fig. 2. For the EasyCom AR glasses, we only used the first 4 microphones which are on the device. Since EasyCom has an audio/video frame rate of 20 Hz, while the frame rate of our M-BEST-RQ encoder is 25 Hz, we resampled the frames in EasyCom by 0.8 through repetition and subsampling.

B. Pre-training

We trained four M-BEST-RQ models on different combinations of datasets: LS, RD, LS+RD, and LL. All models share the same architecture, containing fixed beamformers, a log-mel filterbank extractor, gated 2-D convolutions, and a VGG-conformer encoder. The VGG-Conformer encoder consists of 2 VGG [21] subsampling layers and 24 Conformer encoder layers [22] with a hidden dimension of 512.

[1]Since Libri-Light does not contain transcriptions for each utterance, we used an in-house ASR model to get pseudo-labels for simulation purposes.

[2]https://www.chimechallenge.org/challenges/chime8/task3



(a) Project Aria glasses **(b)** EasyCom AR glasses

Fig. 2: Microphone positions of two devices: (a) Aria glasses, containing 7-channel input, and (b) EasyCom AR glasses, containing 6-channel input. For (b), we only used the first 4 microphones which are on the device.

The number of trainable parameters of M-BEST-RQ is ∼96M. We used 32 NVIDIA A100 GPUs to train M-BEST-RQ on LS, RD, and LS+RD with a batch size of 3000, and 128 A100 GPUs to train M-BEST-RQ on LL with a batch size of 8000. For all pre-training, the Adam optimizer was used with betas of (0.9, 0.98), epsilon of $1e^{-8}$, and weight decay of $1e^{-4}$. A tri-stage learning rate schedule was used with a peak learning rate of 0.0003, warmed up for 20k steps (or 30k for LL pre-training). We used 2048 code-books, each of size 24, for the random quantizer. The mask probability and lengths were set to 0.02 and 30 frames, respectively, based on our preliminary investigation. For fine-tuning, we selected models from different checkpoints for each of LS (600k steps), RD (400k steps), LS+RD (600k steps), and LL (850k steps), based on the convergence of the training accuracy of codebooks.

C. Downstream Tasks

As mentioned earlier, most existing work on multi-channel SSL has conducted evaluations on simulated tasks. In order to evaluate M-BEST-RQ on real settings, we curated three downstream tasks focused on the smart glasses use case. We describe these tasks and their implementation details below.

1) Conversational ASR

The C-ASR task is based on CHiME-8 Task-3, using the MMCSG dataset. It consists of conversations recorded between the wearer and a participant (the speaker located directly in front of the wearer) in the presence of background noise and distracting speakers. The task objective is to transcribe and attribute the speech from both the wearer (self) and the participant (other). Performance is measured using the speaker-attributed WER metric. We used the same data preparation process as the official baseline system for MMCSG. This involves cropping each recording to ∼20 second segments and using serialized output training (SOT) transcripts [23]. We inserted »0 and »1 before each token in the reference to indicate whether the token is attributed to self or other, respectively.

For M-BEST-RQ fine-tuning, we added a 4096-dimensional linear head (one for each sentence-piece including »0 and »1) on the output of the VGG-Conformer encoder and trained with CTC loss [24]. We used 8 A100 GPUs during fine-tuning with batch size 256. We warmed up the learning rate to $3e^{-5}$ for 6000 warm-up steps and then decayed it exponentially. During fine-tuning, we used real volume perturbation to scale the volume of conversations within the range from 0.01 to 0.99. We also added a SpecAugment layer [25] after the feature extraction.

Since the official challenge baseline is a streaming RNN-T model [26], we prepared a comparable ASR baseline which shares the same model architecture as our fine-tuned M-BEST-RQ. We first trained this ASR model on LS, and then fine-tune on MMCSG train set after convergence.

2) Spherical Active Source Localization

The presence of multiple microphones on smart glasses enables localization for the active source around the wearer. S-ASL [27] predicts a (90, 180) feature map where 90 and 180 denote the elevation and azimuthal, respectively, with a 2° resolution. The position in the feature map indicates the angles of directions and is computed by transferring the annotation of 3D points and quaternions to the (90, 180) map where 0/1 indicate the absence/presence of a speech source. Following [28], we treat this task as classification and follow the same step in terms of training and evaluation. During fine-tuning, we added two linear layers at the output of the VGG-Conformer encoder, projecting the hidden dimension to 4050. After reshaping to (45, 90), two tensors are upsampled and concatenated to (2, 90, 180). We apply the same non-maximum suppression of radius 5 and threshold 0 and match the augmented ground truth with the Hungarian algorithm [28]. The evaluation metrics are mean angular errors (MAE) for the distance from prediction to ground truth (indicating false positives), and from ground truth to prediction (indicating missing targets). Training was done on 16 A100 GPUs with a batch size of 64, and a tri-stage learning rate with 2500 warm-up and 17500 decay steps, with a peak learning rate of $3e^{-5}$. Same as [27], [28], we used sessions 4-12 for training and 1-3 for testing on EasyCom.

3) Glasses Wearer VAD

We define this task as frame-level binary classification where for each frame, the model output is 1 if the glasses wearer is speaking and 0 otherwise. We used the same EasyCom dataset as in S-ASL to create this task. Following [27], for the fine-tuning of M-BEST-RQ, we added two linear layers at the end of the VGG-Conformer encoder to perform binary classification. The training was done on 8 A100 GPUs with a batch size 128, with a peak learning rate of $3e^{-5}$ warmed up for 3000 steps. As before, we used sessions 4-12 for training and 1-3 for testing. We computed the mean average precision (mAP) as the metric to evaluate each model.

IV. Results & Discussion

A. Conversational ASR

We compare the results of different systems in Table I. The challenge baseline is an RNN-T ASR model [26] which has ∼114 M trainable parameters, and is pre-trained on 4.5k hours of 7-ch perturbed LS and TED-LIUM [29] (denoted as LS++). We reproduce the official baseline results here [17]. All other systems are models trained with CTC loss with ∼98 M trainable parameters. For the CTC models, we included a system trained directly on the MMCSG training set in addition to our ASR baseline. Systems (A), (B), (C), and (D) use M-BEST-RQ encoders pre-trained on RD, LS, LS+RD, and LL, respectively. We also included systems (E) and (F), which use LS for fine-tuning, in addition to MMCSG.

First, we see that our CTC-based ASR baseline is competitive with the challenge baseline, despite using less pre-training data. Among the M-BEST-RQ systems fine-tuned only with 8h of MMCSG data, systems (A) and (B), trained on RD and LS, respectively, were found to be worse than the ASR baseline. However, system (C), which was trained on LS+RD, achieved 21.5%/29.5% WER on self/other speaker, outperforming the ASR baseline by over 1% absolute WER reduction. This indicates that pre-training using a combination of synthetic and real data may be important for the M-BEST-RQ model, when the size of synthetic data is small. Nevertheless, if large-scale simulated data is used, real data may not be required, as shown by the strong performance of system (D). This system, despite being pre-trained on LL only, achieved 20.1%/28.1% WER on self/other, outperforming the ASR baseline by 2%. Systems (E) and (F) further improved the WER by over 6% absolute, demonstrating the importance of labeled data.

TABLE I: C-ASR results on the MMCSG evaluation set, in terms of self and other WER (%). We also report WER breakdown into insertion (ins), deletion (del), substitution (sub), and speaker attribution (attr) errors. † denotes the official challenge baseline. Underline denotes labeled data in pre-training. All fine-tune data are labeled.

Model	Pre-train		Fine-tune	Self					Other				
	Data	Size (h)	Size (h)	**WER**	ins	del	sub	attr	**WER**	ins	del	sub	attr
RNN-T†	LS++	4.5k	8	22.0	2.7	4.3	13.5	1.6	32.8	4.2	8.0	17.9	2.6
CTC	None	0	8	67.8	4.0	15.2	46.2	2.4	76.4	4.8	14.9	52.1	4.5
	LS	2k	8	22.9	2.5	4.3	14.8	1.4	30.8	3.7	7.3	17.9	1.9
M-BEST-RQ	(A) RD	800	8	40.5	4.0	7.2	28.0	1.4	49.6	5.3	9.4	32.7	2.2
	(B) LS	2k	8	24.3	2.2	4.1	16.6	1.4	33.4	3.8	7.2	20.5	1.8
	(C) LS+RD	2.8k	8	21.5	**2.0**	3.7	14.3	1.4	29.5	**3.5**	6.7	17.9	1.4
	(D) LL	140k	8	**20.1**	**2.0**	**3.5**	**13.4**	**1.2**	**28.1**	3.7	**6.2**	**16.8**	**1.3**
	(E) LS+RD	2.8k	2k+8	16.7	1.8	**3.2**	10.5	**1.1**	24.5	3.3	**6.2**	13.9	**1.1**
	(F) LL	140k	2k+8	**16.5**	**1.6**	3.6	**10.2**	**1.1**	**23.8**	**3.0**	6.7	**12.9**	1.3

TABLE II: S-ASL results on the EasyCom dataset. We report $\text{MAE}_{p\to g}$ (false positives), $\text{MAE}_{g\to p}$ (missing targets), and their mean mMAE, all in degrees. "AV" indicates whether audio-visual input is used in the model. M-BEST-RQ models (C) and (D) are as defined in Table I.

Model (← Input)	AV	Size (M)	$\text{MAE}_{p\to g}$	$\text{MAE}_{g\to p}$	mMAE
[27] ← DOA	✗	15.8	129.8	46.5	88.1
[27] ← AV + cor	✓	28.4	16.8	6.6	11.7
[27] ← AV + spec	✓	28.4	8.8	6.2	7.5
[27] ← DOA + image	✓	28.4	66.8	36.5	51.7
[27] ← AV + raw-audio	✓	28.4	40.1	140.8	90.5
[28] AVSL (scratch)	✓	10.7	9.3	4.7	7.0
[28] AVSL (pre-trained)	✓	10.7	**8.0**	**4.5**	**6.3**
(C) + frozen	✗	4.2	25.9	6.4	16.2
+ weighted comb.	✗	4.2	24.0	4.8	14.4
+ full fine-tune	✗	99.7	4.9	7.0	6.0
(D) + frozen	✗	4.2	26.7	6.3	16.5
+ weighted comb.	✗	4.2	22.0	**4.6**	13.3
+ full fine-tune	✗	99.7	**4.5**	6.7	**5.6**

B. Spherical Active Source Localization

Following [30], we compared the MAE from prediction to ground truth ($p \to g$) and from ground truth to prediction ($g \to p$), and their mean (mMAE), as shown in Table II. We report the baseline numbers directly from the cited papers. In addition to full model fine-tuning, we also evaluated other fine-tuning approaches: (i) *frozen*, where the M-BEST-RQ encoder is kept frozen and only train the last linear layers are fine-tuned, and (ii) "weighted comb.", which additionally uses a weighted combination of all conformer layer outputs with trainable weights. Despite having a much smaller number of trainable parameters, these models outperformed the AV models which use DOA+image and AV+raw-audio as inputs. Nevertheless, the $\text{MAE}_{p\to g}$ was found to be high, indicating that these models are more likely to hallucinate extra speech sources. Full model fine-tuning was able to solve this problem, with the fine-tuned system (D) outperforming all baselines trained on audio-visual (AV) inputs with a state-of-the-art mMAE of 5.6 degrees.

C. Glasses Wearer Voice Activity Detection

Table III shows the mAP numbers for the W-VAD task, with the baselinee numbers reported directly from the cited papers. All baselines were initialized from the AV models trained on the S-ASL task, whereas our models are fine-tuned only on the W-VAD task. We found that our fine-tuned model (C) achieved over 90% mAP, which is comparable with the baselines. Nevertheless, the best baseline results are obtained using spectrogram features, suggesting that the use of log-Mel features in M-BEST-RQ may be suboptimal. Furthermore, since EasyCom does not provide official mouth-directed ATFs, we designed the mouth beamformer solely based on the array geometry. This may also explain the relatively weaker performance, as the beamformed signal from the wearer's mouth may be a strong indicator for the W-VAD task.

TABLE III: W-VAD results on the EasyCom dataset, in terms of mAP (↑). All baselines contain <1M trainable params, similar to the "frozen" and "weighted comb." versions of our M-BEST-RQ models. "AV" indicates whether audio-visual input is used during pre-training.

Model (← Input)	AV	Pre-train	mAP
[27] ← cor		✓	90.20
[27] ← energy		✓	88.89
[27] ← spec		✓	91.69
[27] ← AV + raw-audio		✓	87.29
[28] AVSL		✓	**93.70**
(C) + frozen		✗	86.66
+ weighted comb.		✗	87.75
+ full fine-tune		✗	**90.16**
(D) + frozen		✗	86.12
+ weighted comb.		✗	87.72
+ full fine-tune		✗	89.29

V. Conclusion

We introduced M-BEST-RQ, the first foundation model designed specifically for smart glasses, and curated three downstream tasks to evaluate its performance: C-ASR, S-ASL, and W-VAD. Evaluations on the MMCSG and EasyCom datasets demonstrated the utility of M-BEST-RQ for multi-channel speech across devices. On the conversational ASR task, M-BEST-RQ fine-tuned with only 8 hours labeled data outperformed strong ASR baselines trained on 2k+ labeled hours. With EasyCom, we showed the cross-device generalizability of M-BEST-RQ, where it matched or outperformed state-of-the-art results on the source localization and wearer VAD tasks. We believe that advancements in these tasks have significant potential to improve the user experience in wearable devices. With this framework in place, future work can build upon our model by exploring different SSL techniques or input features, and developing streaming and lightweight foundation models to facilitate seamless deployment.

Acknowledgments. We thank Kateřina Žmolíková, Morrie Doulaty, Christi Miller, and Calvin Murdock for their help in preparing the pre-training data and downstream tasks.

REFERENCES

[1] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al., "Project Aria: A new tool for egocentric multi-modal AI research," *arXiv:2308.13561*, 2023.

[2] Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, et al., "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179–1210, 2022.

[3] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Proc. INTERSPEECH*, 2019, pp. 3465–3469.

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

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

[6] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[7] Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee, "SUPERB: Speech processing universal performance benchmark," in *Proc. INTERSPEECH*, 2021, pp. 1194–1198.

[8] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al., "On the opportunities and risks of foundation models," *arXiv:2108.07258*, 2021.

[9] Antoni Dimitriadis, Siqi Pan, Vidhyasaharan Sethu, and Beena Ahmed, "Spatial HuBERT: Self-supervised spatial speech representation learning for a single talker from multi-channel audio," *arXiv:2310.10922*, 2023.

[10] Qiushi Zhu, Jie Zhang, Yu Gu, Yuchen Hu, and Lirong Dai, "Multichannel AV-wav2vec2: A framework for learning multichannel multi-modal speech representation," in *Proc. of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 19768–19776.

[11] Zili Huang, Yiwen Shao, Shi-Xiong Zhang, and Dong Yu, "UniX-Encoder: A universal x-channel speech encoder for ad-hoc microphone array speech processing," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 11991–11995.

[12] Ju Lin, Niko Moritz, Ruiming Xie, Kaustubh Kalgaonkar, Christian Fuegen, and Frank Seide, "Directional speech recognition for speaker disambiguation and cross-talk suppression," in *Proc. INTERSPEECH*, 2023, pp. 3522–3526.

[13] Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, and Yonghui Wu, "Self-supervised learning with random-projection quantizer for speech recognition," in *Proc. International Conference on Machine Learning*, 2022, pp. 3915–3924.

[14] Ryan Whetten, Titouan Parcollet, Marco Dinarelli, and Yannick Estève, "Open implementation and study of BEST-RQ for speech processing," *arXiv:2405.04296*, 2024.

[15] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 5206–5210.

[16] Jacob Kahn, Morgane Riviere, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al., "Libri-Light: A benchmark for ASR with limited or no supervision," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 7669–7673.

[17] Kateřina Žmolíková, Simone Merello, Kaustubh Kalgaonkar, Ju Lin, Niko Moritz, Pingchuan Ma, Ming Sun, Honglie Chen, Antoine Saliou, Stavros Petridis, Christian Fuegen, and Michael Mandel, "The CHiME-8 MMCSG challenge: Multi-modal conversations in smart glasses," in *Proc. CHiME-8*, 2024.

[18] Jacob Donley, Vladimir Tourbabin, Jung-Suk Lee, Mark Broyles, Hao Jiang, Jie Shen, Maja Pantic, Vamsi Krishna Ithapu, and Ravish Mehra, "EasyCom: An augmented reality dataset to support algorithms for easy communication in noisy environments," *arXiv:2107.04174*, 2021.

[19] Feng-Ju Chang, Martin H. Radfar, Athanasios Mouchtaris, Brian King, and Siegfried Kunzmann, "End-to-end multi-channel transformer for speech recognition," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5884–5888, 2021.

[20] Ju Lin, Niko Moritz, Yiteng Huang, Ruiming Xie, Ming Sun, Christian Fuegen, and Frank Seide, "AGADIR: Towards array-geometry agnostic directional speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 11951–11955.

[21] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. International Conference on Learning Representations*, 2015, pp. 1–14.

[22] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al., "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. INTERSPEECH*, 2020, pp. 5036–5040.

[23] Naoyuki Kanda, Yashesh Gaur, Xiaofei Wang, Zhong Meng, and Takuya Yoshioka, "Serialized output training for end-to-end overlapped speech recognition," in *Proc. INTERSPEECH*, 2020, pp. 2797–2801.

[24] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. International Conference on Machine Learning*, 2006, pp. 369–376.

[25] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. INTERSPEECH*, 2019, pp. 2613–2617.

[26] Vahid Noroozi, Somshubra Majumdar, Ankur Kumar, Jagadeesh Balam, and Boris Ginsburg, "Stateful Conformer with cache-based inference for streaming automatic speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 12041–12045.

[27] Hao Jiang, Calvin Murdock, and Vamsi Krishna Ithapu, "Egocentric deep multi-channel audio-visual active speaker localization," in *Proc. IEEE/CVF Computer Vision and Pattern Recognition Conference*, 2022, pp. 10544–10552.

[28] Jinzheng Zhao, Yong Xu, Xinyuan Qian, and Wenwu Wang, "Audio visual speaker localization from egocentric views," *arXiv:2309.16308*, 2023.

[29] François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve, "TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation," in *Proc. International Conference on Speech and Computer*, 2018, pp. 198–208.

[30] Heeseung Yun, Ruohan Gao, Ishwarya Ananthabhotla, Anurag Kumar, Jacob Donley, Chao Li, Gunhee Kim, Vamsi Krishna Ithapu, and Calvin Murdock, "Spherical world-locking for audio-visual localization in egocentric videos," in *Proc. European Conference on Computer Vision*, 2024, pp. 256–274.