DO YOU LISTEN WITH ONE OR TWO MICROPHONES? A UNIFIED ASR MODEL FOR SINGLE AND MULTI-CHANNEL AUDIO

Gokce Keskin, Minhua Wu, Brian King, Harish Mallidi, Yang Gao, Jasha Droppo, Ariya Rastrow, Roland Maas

Amazon.com

ABSTRACT

Automatic speech recognition (ASR) models are typically designed to operate on a single input data type, e.g. a single or multi-channel audio streamed from a device. This design decision assumes the *primary* input data source does not change and if an additional (*auxiliary*) data source is occasionally available, it cannot be used. An ASR model that operates on both primary and auxiliary data can achieve better accuracy compared to a primary-only solution; and a model that can serve both *primary-only* (PO) and *primary-plus-auxiliary* (PPA) modes is highly desirable. In this work, we propose a unified ASR model that can serve both modes. We demonstrate its efficacy in a realistic scenario where a set of devices typically stream a single primary audio channel, and two additional auxiliary channels *only when* upload bandwidth allows it. The architecture enables a unique methodology that uses both types of input audio during training time. Our proposed approach achieves up to 12.5% relative word-error-rate reduction (WERR) compared to a PO baseline, and up to 16.0% relative WERR in low-SNR conditions. The unique training methodology achieves up to 2.5% relative WERR compared to a PPA baseline.

Index Terms: speech recognition, multi-channel

1. INTRODUCTION

ASR models are typically designed to assume that all input sources to the model are always available for each sample. The inputs could be acoustic data from a single audio channel, from multiple channels, or acoustic data combined with context vector embeddings from prior utterances in a conversational setting [1, 2, 3]. This design choice prevents the ASR model to accept additional input sources that contain useful information but are only *occasionally* available. For instance, consider a classroom scenario where there is a central listener device with a microphone array and an additional lapel microphone that is occasionally used by the teacher. The input to the ASR could be a single audio channel that comes from an on-device beamformer in the central listener. If the upload bandwidth allows it, additional raw microphone channels (auxiliary inputs) could also be streamed. If the lapel is used, another auxiliary source is available. One would need to build separate ASR models for these scenarios.

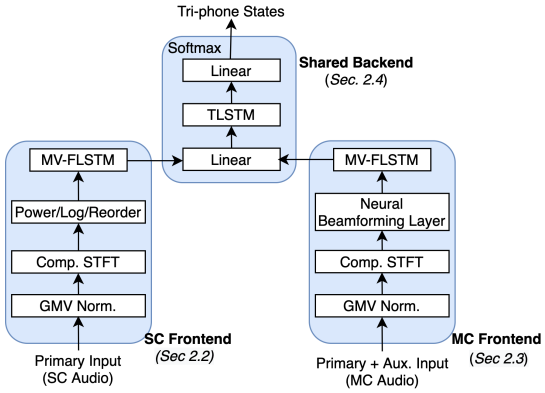

Fig. 1. A Unified ASR model architecture. Separate frontends for each input type (primary and primary+auxiliary) share a backend, enabling a single model that serves both data types.

In this work, we propose a unified model that can serve all these predefined scenarios. The unified model has separate frontends for each scenario (Sec. 3) and employs a unique training methodology that combines datasets with different number of sources (Sec. 5). In the rest of this paper, we present the results of such a unified model in a far-field ASR scenario: a wide variety of devices stream single-channel (SC) audio, but a subset of them might conditionally stream additional raw audio from microphones to create a multi-channel (MC) input. The unified model, coupled with the proposed training methodology, leads to lower WER than building an MC-only model that can only be trained with MC audio (Sec. 6).

2. RELATED WORK

Far-field ASR systems are designed to operate in more challenging acoustic conditions compared to a near-field system where the speaker is close to the microphone. Lower signal-to-noise ratio (SNR) in the microphones reduces the word error rate (WER) of the following ASR system. Increased signal degradation with distance, room reverberation, noise, and background speech contribute to this reduction [4].

A complete distant speech recognition (DSR) system typically consists of distinct components such as a voice ac-

978-1-6654-6867-1/22/$31.00 ©2022 IEEE

tivity detector (VAD), speaker localizer (SL), dereverberator, beamformer and acoustic model [5, 6, 7, 8, 9]. Beamforming techniques take advantage of multiple microphones to enhance the audio signal, and is a key component to improve noise robustness of the DSR. Beamforming could be fixed or adaptive. In comparison to fixed beamforming, adaptive techniques have shown that noise robustness of ASR system can be improved with a dereverberation approach or high-order statistics. However, adaptive techniques rely on accurate VAD or SL, and they can underperform in comparison to a fixed beamformer when these dependent components are not performing reliably. Previous studies show individually optimizing various DSR components is sub-optimal [10, 11].

More recently, multi-channel deep neural network (MC-DNN) approaches have been applied to ASR by training a unified MC-DNN model where the MC processing modules are part of the DNN structure [12, 13, 14]. Aside from unified MC-DNN approaches, a DNN is also employed to construct a clean speech signal. A mask-based method was proposed to estimate the statistics of the target clean speech via an LSTM [15, 16]. However, this method needs accumulated statistics from adequate amount of adaptation data to perform well. Accumulating the statistics might cause additional latency and is less applicable to real-time applications.

3. MODEL ARCHITECTURE

3.1. Overview

The unified architecture diagram is given in Fig. 1. The model includes two separate frontends for SC and MC audio, with a shared backend. SC audio, or the *primary channel*, is either obtained from the only existing microphone in the device or from an on-device beamformer that combines the outputs of multiple microphones. In our case, MC audio has three channels: the primary channel and two *auxiliary channels* that are obtained from the raw audio outputs of two microphones. The model is trained with a mix of SC and MC audio. MC samples propagate through the MC frontend (FE) and the shared backend, whereas SC samples propagate through the SC FE and the shared backend. During inference, audio is propagated through the corresponding FE for the incoming data type, followed by the shared backend (Sec. 5).

3.2. SC Frontend

In the SC FE, extracted audio features are processed by a multi-view frequency LSTM (MV-FLSTM) [17], followed by a shared backend (Sec. 3.4). For feature extraction, global mean and variance (GMV) values computed from the received channel are used to normalize the input. Complex STFT features with 256 frequency bins are extracted from the normalized waveform with a window size of 25ms produced in 10ms steps. Three consecutive input frames are stacked into a single vector to reduce the number of time steps in the

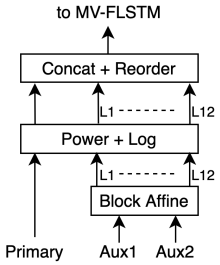

Fig. 2. Neural Beamforming Layer.

acoustic model, known as the Lower Frame Rate model [18]. MV-FLSTM uses the log-power of the complex features. Frequency LSTMs operate over the frequencies contained in an individual input frame with a sliding window [19]. MV-FLSTM extends this concept by having several window sizes (*views*) to span different frequency ranges at each step. For details, we refer readers to prior work which has shown 3%-7% relative gains in WER by using the MV-FLSTM approach over a single-view FLSTM [17].

3.3. MC Frontend

MC audio is normalized by GMV statistics for each channel. Neural Beamforming Layer (NBL) combines the audio channels and produces a number of look directions, which are processed by a separate MV-FLSTM and the shared backend. NBL implementation is adopted from the Elastic Spatial Filter described in [2]. The Discrete Fourier Transform of the normalized input signal can be described as below:

$$\mathbf{X}\left(t,\omega_k\right) = \left[X_1\left(t,\omega_k\right),\cdots,X_M\left(t,\omega_k\right)\right]^T \quad (1)$$

Using this notation, we can express the complex weight vector for source position $\mathbf{p}$ as follows:

$$\mathbf{w}^H\left(t,\omega_k,\mathbf{p}\right) = \left[w_1\left(t,\omega_k,\mathbf{p}\right),\cdots,w_M\left(t,\omega_k,\mathbf{p}\right)\right] \quad (2)$$

Thus, the block affine transform (BAT) can be expressed as:

$$\begin{bmatrix} Y_1\left(\omega_1\right) \\ \cdots \\ Y_D\left(\omega_1\right) \\ \cdots \\ Y_1\left(\omega_K\right) \\ \cdots \\ Y_D\left(\omega_K\right) \end{bmatrix} = \begin{bmatrix} \mathbf{w}_{\mathrm{SD}}^H\left(\omega_1,\mathbf{p}_1\right)\mathbf{X}\left(\omega_1\right)+\mathbf{b}_1 \\ \cdots \\ \mathbf{w}_{\mathrm{SD}}^H\left(\omega_1,\mathbf{p}_D\right)\mathbf{X}\left(\omega_1\right)+\mathbf{b}_D \\ \cdots \\ \mathbf{w}_{\mathrm{SD}}^H\left(\omega_K,\mathbf{p}_1\right)\mathbf{X}\left(\omega_K\right)+\mathbf{b}_{D(K-1)+1} \\ \cdots \\ \mathbf{w}_{\mathrm{SD}}^H\left(\omega_K,\mathbf{p}_D\right)\mathbf{X}\left(\omega_K\right)+\mathbf{b}_{DK} \end{bmatrix} \quad (3)$$

where $\mathbf{b}$ is bias term, D is the number of look directions and K is the number of frequency bins.

Complex STFT features from the two microphones are processed by the trainable BAT, generating 12 look directions (Fig. 2). BAT is initialized with super directive beamformer weights. Log-power features of the look directions and the primary channel are concatenated and further processed by the MC FE's MV-FLSTM. Early experiments showed that using all three channels in the MC FE had 4-7% relative WERR

compared to using only the two auxiliary (raw) channels, hence the primary input is concetanated to the BAT output. MC FE's MV-FLSTM similar to the SC FE's MV-FLSTM; the main difference is the 13X larger input dimension. This is due to the concatenation of the primary channel and the look directions from the NBL.

3.4. Shared Backend

The shared backend has a projection (linear) layer to reduce the input dimensionality to the five-layer, unidirectional TLSTM with 768 cells per layer. Unidirectionality is required to preserve causality for a streamable ASR model. TLSTM output is connected to a classification layer with softmax outputs to generate tied tri-phone states used in a hybrid ASR model [20]. Unified model has a total of 28M parameters.

4. DATASETS

Two mutually-exclusive training datasets are used in this work: SC datasets [$D500_{sc}$, $D10K_{sc}$, $D65K_{sc}$] are human transcribed with [500, 10K, 65K] hours of audio and they contain only the primary channel. MC datasets [$D500_{mc}$, $D10K_{mc}$, $D20K_{mc}$] have [500, 10K, 20K] hours of audio respectively and contain three channels: primary channel and two auxiliary channels obtained from raw audio of two microphones. MC datasets are transcribed by a full-context SC model with relaxed constraints on latency, causality and size. $DTest_{mc}$ is a human-transcribed MC test set with 45 hours of audio, containing three channels (one primary and two auxiliary). For evaluation of SC models, only the primary channel in this set is used. The test set contains a mix of single and multi-speaker utterances (i.e., background speech). SNR value for each utterance is also available in the test set to evaluate performance across different noise conditions. All audio data is de-identified for privacy reasons.

5. TRAINING METHODOLOGY

During training, each batch of data contains a mix of MC and SC audio. Since the primary channel is present in the MC dataset, the auxiliary channels can be removed and the resulting SC data is added to expand the SC dataset. Gradient updates obtained from the expanded SC dataset are used to train the SC FE and the shared backend. Gradients generated from the MC data are used to update the MC FE and the shared backend. The unified ASR architecture described (Sec. 3) allows the shared backend to learn from both SC and MC data.

There are other alternatives to incorporate SC data to MC model training. One can pad the missing two channels in the SC data with zeros to expand the MC dataset and train an MC model with this expanded dataset. However, experiments show this is inferior to unified model training with two separate frontends (Sec. 6.1). Another alternative is to first train an SC-only model to obtain the backend, freeze its weights, append an MC FE and then train the MC FE with MC-only data. Freezing the weights is required to enable a single model for SC and MC audio. This approach is undesirable since it requires a two-step methodology and the backend will not be updated for MC data. In practice, training in this manner led to non-convergence of the MC FE in large-scale datasets.

Table 1. Normalized WER Comparison of Unified Model with Two Frontends to Zero-Padding Missing Channels.

Exp.	Train Data	Model / Inf. Path	nWER (Single/Multi-Speaker)
E1	$D500_{sc}$ +$D500_{mc}$	MC, Zero-pad missing channels	100.0/100.0
E2	$D500_{sc}$ +$D500_{mc}$	Unified (MC FE + Shared Backend)	**96.0/96.9**

6. EXPERIMENTAL RESULTS

Models are trained with cross-entropy (CE) loss, followed by CTC loss [21] for the same number of training epochs as the corresponding baseline. WER results are obtained from the test set $DTest_{mc}$ (Sec. 4) and normalized to the baseline. Absolute WER values are below 10% for medium and large-scale datasets. Only the primary channel is used for SC-only models during training and test.

6.1. Comparison of Proposed Approach to Zero Padding

Table 1 shows the Normalized WER (nWER) of the zero-padding baseline (E1) to the proposed unified model architecture (E2). E1 is a combined MC/SC model that consists of the same MC FE and the shared backend as E2 (Sec. 3), but does not include an SC FE. E1 always expects three input channels (one primary and two auxiliary); if only SC data is available during training or inference, the missing channels are filled with zeros. E2 is trained with two separate frontends (Sec. 5), with each input type going through its respective frontend. During test time, all three channels in $DTest_{mc}$ are used as input to E1 and E2. WER values of E1 for single and multi-speaker cases are arbitrarily set to 100.0, and E2's WER values are normalized to E1. The unified model architecture achieves 3.1%-4.0% relative WERR compared to the baseline, demonstrating its advantage over the alternative.

6.2. Impact of Including Multi-Channel Audio

Table 2 shows the nWER comparison of the SC and MC models in medium and large-scale datasets. E3 has the same SC FE and the shared backend architecture in Fig. 1, but does not contain the MC FE. Datasets $D10K_{sc}$ and $D10K_{mc}$ are combined for training (Sec. 4); however E3 only uses the primary

Table 2. Normalized WER Comparison of MC and SC models.

Exp.	Training Data	Model / Inference Path	nWER (Single/Multi-Speaker)
E3	$D10K_{sc}$ + $D10K_{mc}$	Standalone SC	100.0/100.0
E4	$D10K_{mc}$	Standalone MC	89.7/94.8
E5	$D10K_{sc}$ + $D10K_{mc}$	Unified (SC FE + Shared Backend)	101.5/99.3
E6	$D10K_{sc}$ + $D10K_{mc}$	Unified (MC FE + Shared Backend)	**87.5/92.3**
E7	$D65K_{sc}$ + $D20K_{mc}$	Unified (SC FE + Shared Backend)	85.9/88.7
E8	$D65K_{sc}$ + $D20K_{mc}$	Unified (MC FE + Shared Backend)	**82.6/85.7**

channel in $D10K_{mc}$ since it is an SC-only model. WER values of E3 for single and multi-speaker cases are arbitrarily set to 100.0, and following experiments' WER values are normalized to E3. E4 is an MC model that combines the MC FE and the shared backend, and no SC FE. It is trained with $D10K_{mc}$ and uses all three channels. E3 and E4 are trained separately, with no weight sharing. E4 obtains a 10.3% relative WER reduction compared to the SC model (E3) in the single-speaker test set, even though E4 is trained with half the data of E3. This clearly demonstrates the improvements that can be obtained from using multiple audio channels. Multi-speaker test set has 5.2% reduction.

6.3. Impact of Unified Model Training

Experiment E5 shows the results for the SC path (SC FE + shared backend) in the unified model and E6 shows the MC path results for the same model. E5 and E6 are trained together (Sec. 5). Compared to E4, E6 achieves a further 2.2%/2.5% WERR in single/multi-speaker conditions. This demonstrates an advantage of the proposed architecture. E5 has a 0.7% improvement in multi-speaker over E3, but a 1.5% degradation in single-speaker conditions is observed. It is not clear if this degradation is an artifact of the unified model methodology or due to the inherent variability in training dynamics. Multiple runs with different initialization seeds might be warranted for further study. Table 2 also shows the nWER comparison of the unified model with SC (E7) and MC (E8) paths in a large-scale data setting. E7 obtains 14.1%/11.3% reduction in relative WER compared to the SC baseline model E3. E8 has an additional 3.3%/3.0% relative gain over E7. SC data is more widely available due to the prevalence of SC ASR models, and our dataset distribution reflects this fact. Additional experiments with a different data distribution (e.g., oversampling the MC data to reach a 50/50 distribution) could determine if MC results can be further improved.

6.4. Impact of Utterance SNR

Fig. 3 shows the nWER with respect to SNR of utterances. The test set is split into three bins according to utterances' SNR, and WER is computed for each bin. WER for each bin is normalized to the corresponding SNR bin of E3, whose WER is arbitrarily set to 100.0. Comparing the MC model

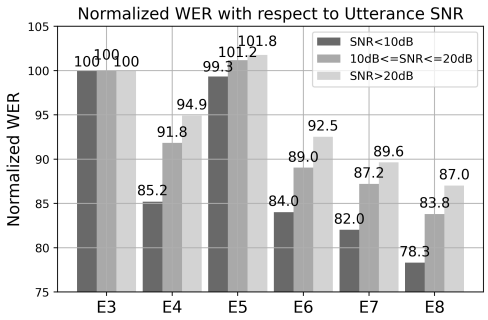

Fig. 3. Relative WERR is more pronounced in low-SNR conditions when an MC model is used (e.g., E4 vs E3). Additional data helps more in low-SNR conditions (E7 vs. E5).

E4 to the SC model E3, the advantage of the MC FE is more pronounced in low-SNR ($< 10dB$) conditions. A significant 14.8% relative WERR is observed in this regime. This is perhaps not surprising since additional information available in the auxiliary channels is even more valuable in these noisy utterances. Conversely, the advantage of incorporating SC data in E6 is more evident in medium and high-SNR conditions, with 2.8% and 2.4% relative WERR compared to E4. In low-SNR conditions, a smaller 1.2% improvement is observed. E6 achieves 16.0% relative WERR in low-SNR conditions compared to the SC baseline E3. Adding large-scale data also helps significantly in low-SNR conditions. Comparing SC-models E5 and E7, a 17.3% relative WERR is observed. In medium-SNR ($[10dB, 20dB]$) and high-SNR ($> 20dB$) conditions, still significant reductions of 14.0% and 12.2% are seen. The combination of all three techniques (unified model, MC FE and additional data) leads to an impressive 21.7% relative WERR in low-SNR conditions (E3 vs. E8), with 16.2% and 13.0% reductions in medium and high-SNR utterances.

7. CONCLUSION

We propose a unified MC/SC model that can be trained with both input types, allowing a single model to support a variety of scenarios. Proposed approach achieves up to 2.5% relative WERR compared to the MC baseline and up to 16.0% relative WERR compared to the SC baseline in low-SNR conditions.

8. REFERENCES

[1] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang, Q. Liang, D. Bhatia, Y. Shangguan, B. Li, G. Pundak, K. C. Sim, T. Bagby, S. Chang, K. Rao, and A. Gruenstein, "Streaming end-to-end speech recognition for mobile devices," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6381–6385.

[2] M. Wu, K. Kumatani, S. Sundaram, N. Ström, and B. Hoffmeister, "Frequency domain multi-channel acoustic modeling for distant speech recognition," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6640–6644.

[3] S. Kim, S. Dalmia, and F. Metze, "Cross-Attention End-to-End ASR for Two-Party Conversations," in *Proc. Interspeech 2019*, 2019, pp. 4380–4384. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-3173

[4] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix, and T. Nakatani, "Far-field automatic speech recognition," *Proceedings of the IEEE*, vol. 109, no. 2, pp. 124–148, 2021.

[5] M. Omologo, M. Matassoni, and P. Svaizer, *Speech Recognition with Microphone Arrays*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 331–353.

[6] M. Wölfel and J. McDonough, *Distant Speech Recognition*. London: Wiley, 2009.

[7] K. Kumatani, T. Arakawa, K. Yamamoto, J. W. McDonough, B. Raj, R. Singh, and I. Tashev, "Microphone array processing for distant speech recognition: Towards real-world deployment," in *Proc. APSIPA ASC*, 2012.

[8] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Häb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP J. Adv. Sig. Proc.*, p. 7, 2016.

[9] T. Virtanen, R. Singh, and B. Raj, *Techniques for Noise Robustness in Automatic Speech Recognition*. West Sussex, UK: John Wiley & Sons, 2012.

[10] J. McDonough and M. Wölfel, "Distant speech recognition: Bridging the gaps," in *Proc. HSCMA*, 2008.

[11] M. L. Seltzer, "Bridging the gap: Towards a unified framework for hands-free speech recognition using microphone arrays," in *Proc. HSCMA*, 2008.

[12] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, "Deep beamforming networks for multi-channel speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5745–5749.

[13] T. Ochiai, S. Watanabe, T. Hori, and J. R. Hershey, "Multichannel end-to-end speech recognition," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 2632–2641.

[14] M. Wu, K. Kumatani, S. Sundaram, N. Ström, and B. Hoffmeister, "Frequency domain multi-channel acoustic modeling for distant speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6640–6644.

[15] J. Heymann, M. Bacchiani, and T. N. Sainath, "Performance of mask based statistical beamforming in a smart home scenario," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6722–6726.

[16] T. Higuchi, K. Kinoshita, N. Ito, S. Karita, and T. Nakatani, "Frame-by-frame closed-form update for mask-based adaptive mvdr beamforming," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 531–535.

[17] M. Van Segbroeck, H. Mallidih, B. King, I.-F. Chen, G. Chadha, and R. Maas, "Multi-view Frequency LSTM: An Efficient Frontend for Automatic Speech Recognition," *arXiv e-prints*, p. arXiv:2007.00131, Jun. 2020.

[18] G. Pundak and T. Sainath, "Lower frame rate neural network acoustic models," in *Interspeech*, 2016.

[19] J. Li, A. rahman Mohamed, G. Zweig, and Y. Gong, "Lstm time and frequency recurrence for automatic speech recognition," *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 187–191, 2015.

[20] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*. Springer Publishing Company, Incorporated, 2014.

[21] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06, 2006, p. 369–376. [Online]. Available: https://doi.org/10.1145/1143844.1143891