Appendix

A ANNOTATE PRONUNCIATION ERRORS IN DATASET

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

Table 1: Annotation for differnet types of pronunciation error.

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

B IMPLEMENTATION DETAILS

B.1 wav2vec 2.0

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

B.2 AV-HuBERT

We have included a concise description of the AV-HuBERT structure in Section 4.3 Video Encoder. The original structure is able to accept both audio and video modalities [8]. Since we only need AV-HuBERT to extract the visual representations, we remove layers handling audio inputs. The video clips are firstly fed into a 3D convolutional frontend. The 3D convolutional frontend has a 3D convolution with $5 \times 7 \times 7$ kernel size and $1 \times 2 \times 2$ stride, a batch normalization [5], a ReLU activation and a 3D max pooling with $1 \times 3 \times 3$ kernel size and $1 \times 2 \times 2$ stride. Then a modified ResNet-18 [3] is adopted to extract the latent visual features. To compensate for the absence of audio modality, we concatenate the latent visual features with zero tensors and remove the modality dropout in the original structure. The fused features are fed into the context network to obtain the final output, which has been elaborated in Section 4.3.

B.3 IMU CRNN

In IMU CRNN (see section 4.4 IMU Encoder), the 1D convolutions have {128, 200} channels for each layer, kernel size of 3, and stride of 1. The strides for max-pooling layers are 2. ReLU activation is applied after each layer. Dropout [9] is applied after each convolution and GRU layer with rates of 0.5 and 0.2, respectively. The GRU contains two layers with 60 hidden units in each layer.

B.4 Experiments

For the Voice Activity Detection (VAD) task, we train the model for 50 epochs, evaluate it at the end of each epoch, and select the model with the lowest validation loss. The model is trained using AdamW optimizer [7], with 1×10^{-3} learning rate.

For the Automatic Lyric Transcription (ALT) task, we train the model for 20 epochs using Adam optimizer [6]. Since we adopt the transfer learning paradigm for wav2vec 2.0 and AV-HuBERT, a small learning rate with 1×10^{-5} is used to prevent catastrophic forgetting. For the rest parts of the ALT system, the learning rate ranges from 1×10^{-5} to 5×10^{-4}.

C MORE QUALITATIVE RESULTS

More qualitative results are displayed in Table 2. It is noticed that our MM-ALT system performs better than its audio-only and audio-visual counterparts by reducing word errors.

REFERENCES

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

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

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

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

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

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

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

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

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

Table 2: More Qualitative Results. Deletions are marked in (brackets and red), and substitutions are marked in *italics and cyan*.

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