Supplementary Material for VLRASN-112: A VLR Dataset for Assamese Compound Numeric Sequences

Meghali Deka$^{[0009-0001-6227-7836]}$, Vaibhav Gavit$^{[0009-0001-6658-7436]}$, Prithwijit Guha$^{[0000-0003-2885-0026]}$, Sukumar Nandi$^{[0000-0002-5869-1057]}$, and Priyankoo Sarmah$^{[0000--0002-9051-1255]}$

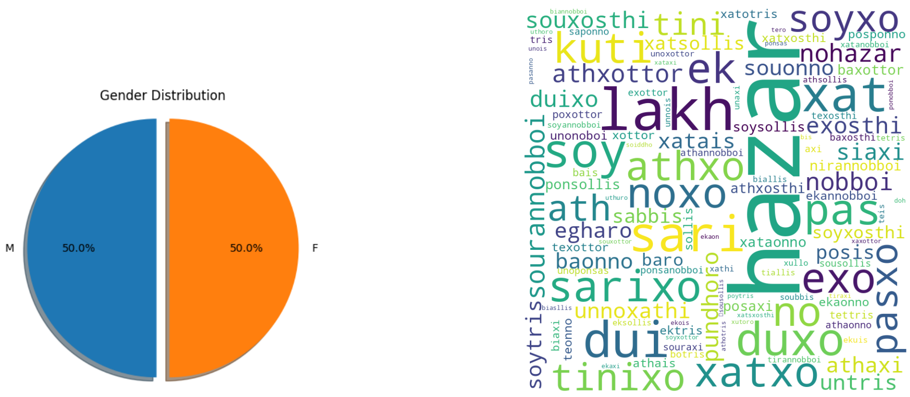
Indian Institute of Technology Guwahati, Assam
{meghali_deka,v.gavit,pguha,sukumar,priyankoo}@iitg.ac.in

1 The VLRASN-112 Dataset: Recording Protocols

- **Frame Rate**: Each video is captured at a frame rate of 30 fps, which ensures a balance between capturing detailed lip movements and managing file size.
- **Camera Positioning**: The camera is positioned to capture a frontal view of the speaker, ensuring a direct line of sight to the speakers face for optimal lip reading.
- **Background**: Both plain and uncluttered backgrounds are chosen for all recordings.
- **Lighting Conditions**: Videos are recorded under various conditions, with a focus on minimizing shadows that may hinder the clarity of the speaker's lip movements.
- **Audio Recordings**: The built-in microphone of the recording device is used to capture clear and natural speech, without the use of external microphones or headphones.
- **Speaking Speed**: Speakers are directed to speak naturally, thereby capturing the difference in varying speaking speeds.
- **Sentence Duration**: The duration of a sentence ranges from a minimum of 2 seconds to 7 seconds.
- **Quantity and Repetition**: Each speaker utters 50 unique sentences, with each sentence being spoken twice.
- **File Naming and Organization**: A systematic naming format is employed:
  `<SPKID>_<SENTENCEID>_<SENTENCENAME>_UTTERANCE_NO.`, facilitating easy data management.
- **Sentence IDs**: Sentences are categorized into sets based on their ID ranges, ensuring organized data segmentation.
  - SET 1 contains sentence ID from 0001 to 0050.
  - SET 2 contains sentence ID from 0051 to 0100.
- **Sentence Name in Assamese**: The transliterated Assamese script is used for naming the sentences.

- **Camera Stability**: The camera is fixed using a tripod stand to prevent any video instabilities.
- **Speaker Frame**: The recording frame includes the speaker from head to shoulder, providing a clear view of the facial features relevant to lip reading.
- **Recording Mode**: All recordings are made in landscape mode.

To create the VLRASN-112 dataset, 30 volunteers were recruited, 15 of whom were female and 15 of whom were male. The distribution is shown in Figure 1(a). This dataset includes 112 unique classes. The word cloud of the dataset is depicted in Figure 1(b).



(a) (b)

Figure 1: (a) Gender statistics across the VLRASN-112 Dataset and (b) Word cloud of the numbers across the VLRASN-112 Dataset

2 Gabor Filters

Gabor filters play a pivotal role in our methodology, given their efficacy in texture analysis and feature extraction.

The Gabor filter [2] is defined by the following equation in the spatial domain:

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \cos\left(\frac{2\pi x'}{\lambda} + \psi\right) \tag{1}$$

$$\text{where, } x' = x\cos\theta + y\sin\theta \text{ and } y' = -x\sin\theta + y\cos\theta$$

Here, σ is the standard deviation of the Gaussian envelope, influencing how broadly information is integrated. Higher values result in broader coverage, enhancing general feature visibility. θ is the orientation of the normal to the parallel stripes of the filterwith vertical (0 degrees) orientations emphasizing facial outlines and horizontal (90 degrees) ones enhancing details like eyes and mouth. λ is the wavelength of the sinusoidal factor.Decreasing wavelength sharpens the filter pattern, suitable for detailed texture analysis, whereas increasing it can enhance general feature recognition. γ is the spatial aspect ratio that modifies the ellipticity of the Gaussian function, vital for processing features with non-uniform spatial distributions, such as those found in facial recognition applications. The kernel size determines the area covered by each filter, with smaller kernels emphasizing detail and larger ones highlighting broader features. By manipulating these parameters, we tailor the Gabor filters to highlight specific features or patterns more effectively, thereby enhancing the accuracy and efficiency of our image processing tasks.

3 Evaluation Metric

The performance of the proposed model is evaluated using word error rate (WER) and character error rate (CER) metrics.

The CER and WER are defined as [1]

$$\text{ErrorRate} = \frac{S + D + I}{N} \tag{2}$$

Where:

S is the number of substitutions,

D is the number of deletions,

I is the number of insertions to get from the reference to the hypothesis, and

N is the number of words in the reference.

WER measures the percentage of incorrectly predicted words compared to ground truth, while CER measures the percentage of incorrectly predicted characters. Additionally, the loss and WER are monitored during training to track model convergence and performance. Lower values of WER and CER indicate better performance in accurately transcribing spoken words.

References

1. Assael, Y.M., Shillingford, B., Whiteson, S., De Freitas, N.: Lipnet: End-to-end sentence-level lipreading. arXiv preprint arXiv:1611.01599 (2016)
2. Imamura, A., Arizumi, N.: Gabor filter incorporated cnn for compression. In: 2021 36th International Conference on Image and Vision Computing New Zealand (IVCNZ). pp. 1–5. IEEE (2021)