Speech intelligibility prediction based on syllable tokenizer

Szymon Drgas

Institute of Automatic Control and Robotics
Poznan University of Technology
szymon.drgas@put.poznan.pl

Abstract

In this report, an intrusive system for speech intelligibility prediction is described. It is based on a pre-trained SD-HuBERT, a neural network that transforms a speech signal to a sequence of embeddings that correspond to syllable-like segments. I propose a neural network that compares such sequences of embeddings using a bilinear neural network architecture. The experimental results show that the proposed system outperforms the baseline HASPI for the CPC3 development data set. Furthermore, after adding internal HASPI features to the proposed system, further improvement is achieved.

Index Terms: speech intelligibility prediction, SD-HuBERT

1. Introduction

In recent work, it was shown that by applying a specific neural network architecture and training procedure, a syllabic organization emerges. This was achieved by fine-tuning the pre-trained HuBERT [1] model with a sentence-level self-distillation method: Self-Distilled HuBERT (SD-HuBERT) [2]. Syllables are large enough to carry robust acoustic cues, yet small enough to map directly onto lexical and phonological representations. Therefore, organizing the speech signal into syllables supports perception.

More technically, SD-HuBERT extracts a sequence of feature vectors which within a syllable are similar (in terms of inner product). This property is used to segment the speech signal into syllable-like segments. Finally, the network outputs a sequence of feature vectors, one vector per syllable, that can be named syllable embeddings.

This report describes a system that uses syllable embedding as input features. I propose a neural network module that performs trainable bilinear comparison of syllable embeddings from the reference signal to feature vectors extracted from the processed signal in time regions corresponding to the mentioned syllable embeddings.

2. Model

In the proposed method, the signals for the left and right ears are independently processed by the so-called bilinear comparison module (BCM) which is described in the next subsection. BCM accepts reference and processed signals together with a value corresponding to the severity of hearing impairment and outputs a speech intelligibility score. The scores for the left and right ears are combined as described in Section 2.2.

2.1. Bilinear comparison module

The proposed module is shown in Figure 1. It combines the features extracted using SD-HuBERT and HASPI [3]. SD-HuBERT extracts a sequence of syllable embeddings from the reference signal, as well as syllable boundaries. The SD-HuBERT features are also extracted from the processed signal (with a rate of 50 feature vectors per second). The boundaries are used to group (average) the features of the processed signals within the syllable regions.

For each ear, two L length sequences of vectors are compared: $\bar{\mathbf{r}}_k$ and $\bar{\mathbf{p}}_k$. First, vectors are normalized using the 2-norm to $\mathbf{r}_k$ and $\mathbf{p}_k$, respectively. Next, H heads specified by trainable diagonal matrices $\mathbf{D}$ are used to transform the $\mathbf{r}_k$ and $\mathbf{p}_k$ to a sequence of score vectors $\mathbf{s}_k$. The h'th component of vector $\mathbf{s}_k$ is calculated as

$$(\mathbf{s}_k)_h = \mathbf{t}_k^{\mathrm{T}} \mathbf{D}_h \mathbf{p}_k$$

Next, multi-head attention is used to transform $\mathbf{s}_k$ to $\mathbf{a}_k$, where query, key, and value are the same matrices. After that, the output of the multi-head attention module is averaged along the temporal index, resulting in a vector of H dimensional (containing four 4-dimensional subvectors corresponding to heads of the multi-head attention module. The resulting vector was concatenated with an 11-dimensional vector representing modulation frequencies of the cepstral coefficients (the vector in HASPI which is provided at the input of its built-in neural network). This vector is processed by a multilayer perceptron (MLP) that outputs the intelligibility score.

2.2. Combination for ears

The combination module accepts two scalar values that represent the predicted speech intelligibility of both ears. These values are combined using the smooth max function with $\tau = 0.1$:

$$o = \tau \log(\exp(i_l/\tau) + \exp(i_r/\tau)) \ ,$$

where i_l and i_r are intelligibility scores for the left and right ears, respectively. Additionally, the hearing loss severity is used, to choose an appropriate trained value, that is used to modulate the combined intelligibility using score FiLM [4]. Finally, a sigmoid layer is applied to get a value between 0 and 1.

3. Experimental setup

3.1. Configuration of the model

The feature vectors of SD-HuBERT are 786-dimensional. The number of bilinear heads was set to 16. The number of attention heads in the MHA module was set to 4. The MLP module accepts concatenation of two vectors: the result of MHA with pooling (16 dimensional) and the vector from HASPI (11 dimensional). It is processed by a fully-connected layer with 54 outputs and ReLU nonlinearity. The second layer has one output.

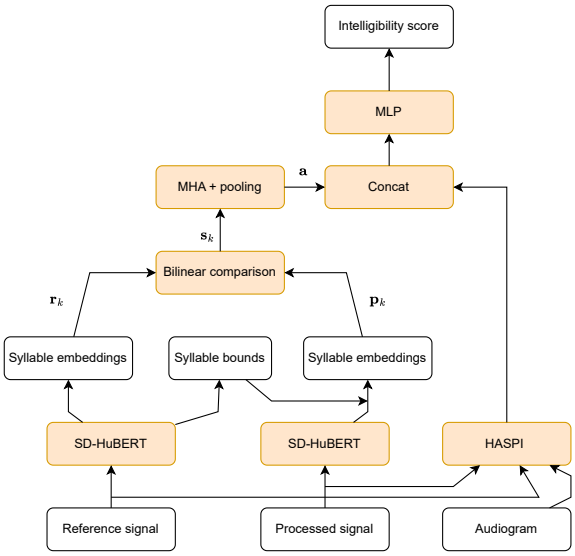

Figure 1: *Bilinear comparison module*

3.2. Configuration of training

In order to train a neural network, three-fold cross-validation was used (stratified according to the hearing loss severity labels). The optimization algorithm chosen was AdamW. The bach size was 16 and the learning rate was set to 0.0001, it was reduced on plateu with patience parameter set to 3. The number of training folds was 30.

4. Results

The results on the development dataset are presented in the table below: It can be noticed that using SD-HuBERT features

System	RMSE	Corr
HASPI	28.0	0.72
BCM	26.73	0.75
BCM with HASPI features	25.61	0.77

Table 1: *The results for the development dataset*

can result with an improvement, both in terms of RMSE and correlation, over the baseline HASPI system.

5. Conclusions

The use of SD-HuBERT features can be used for speech intelligibility prediction. In this report, it was shown that they can outperform HASPI predictions, both in terms of RMSE and correlation.

6. References

[1] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. 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.

[2] C. J. Cho, A. Mohamed, S.-W. Li, A. W. Black, and G. K. Anumanchipalli, "Sd-hubert: Sentence-level self-distillation induces syllabic organization in hubert," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 076–12 080.

[3] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (haspi) version 2," *Speech Communication*, vol. 131, pp. 35–46, 2021.

[4] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.