Orthogonality and Isotropy of Speaker and Phonetic Information in self-supervised speech representations

School of Informatics

Mukhtar Mohamed, Oli Liu, Hao Tang, Sharon Goldwater

University of Edinburgh

NLP UKRI CENTRE FOR DOCTORAL TRAINING

Previous work suggests geometric properties of a representation space reflects its quality.

Orthogonality between phone and speaker subspaces supports simple disentanglement (Liu et al., 2023).

Isotropy in a representation space implies all dimensions are utilized uniformly, which proves helpful in some tasks (e.g. modeling semantic similarity), but harmful in others (e.g. clustering).

In this work, we propose a quantitative measure, *Cumulative Residual Variance*, to evaluate:

Questions

- To what extent do different SSL models exhibit these two geometric properties?
- How do these properties relate to performance on phone and speaker classification?

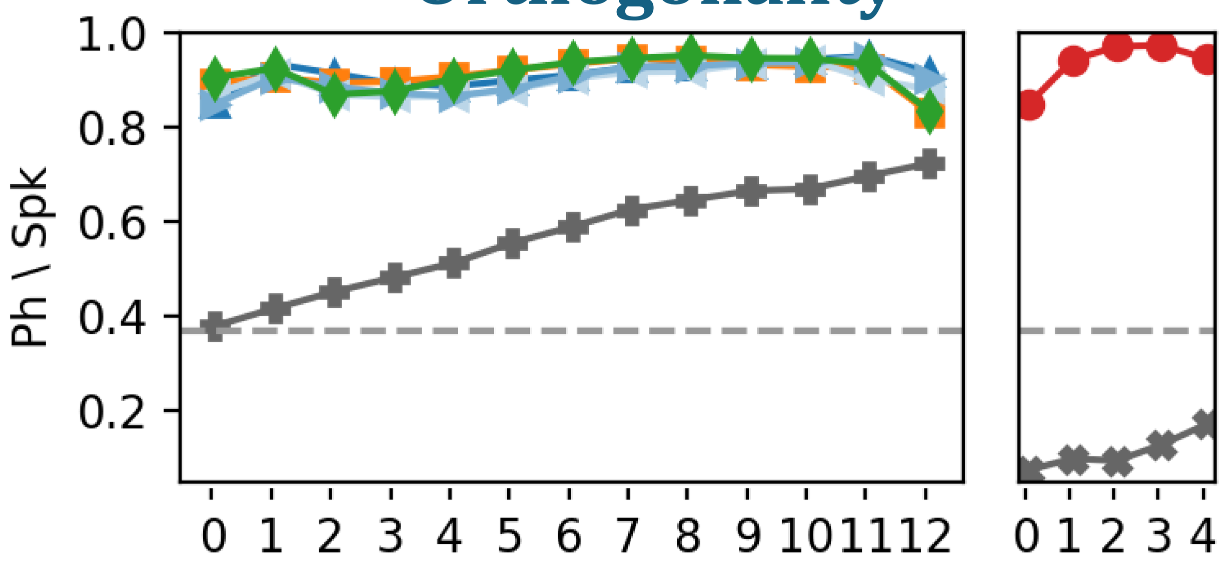
Results

We compared six self-supervised speech models ... ... with untrained models

▲ HuBERT ▶ WavLM ◀ WavLM+ ■ wav2vec 2.0 ◆ data2vec ● CPC-big - - - log Mel ✚ rand. HuBERT ✖ rand. CPC-big

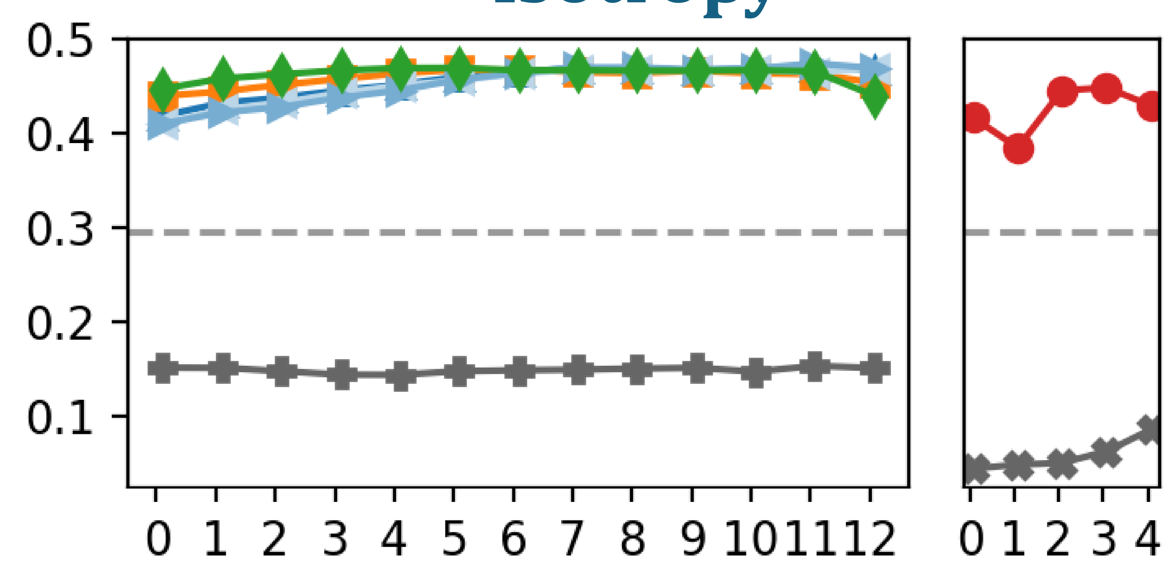
... across all layers within each model ... with acoustic features

Orthogonality



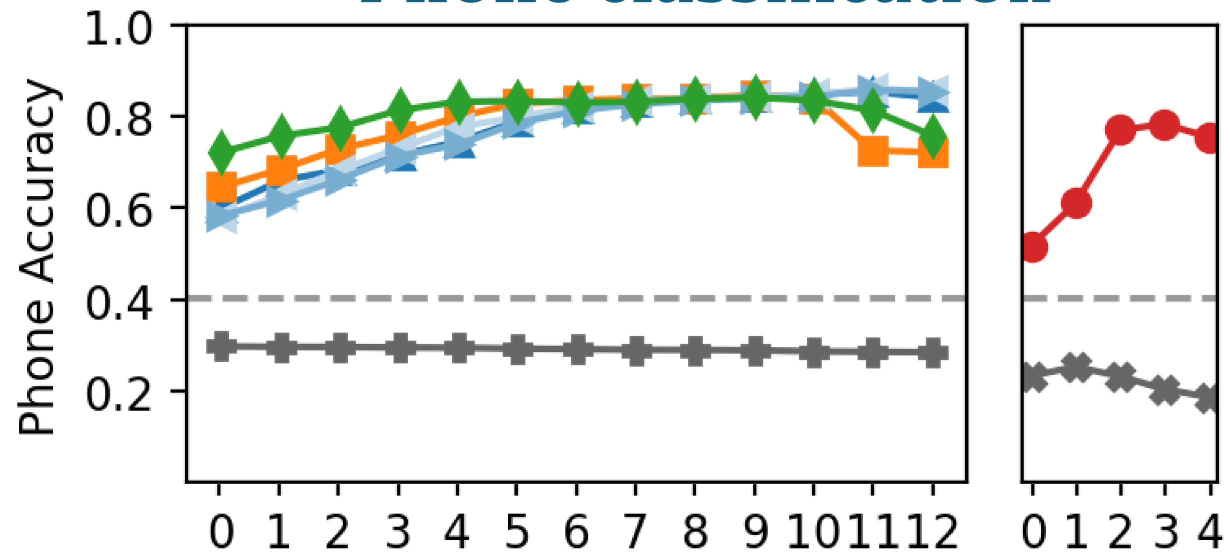
There is a significant degree of orthogonality and isotropy in all 6 trained models.

Isotropy

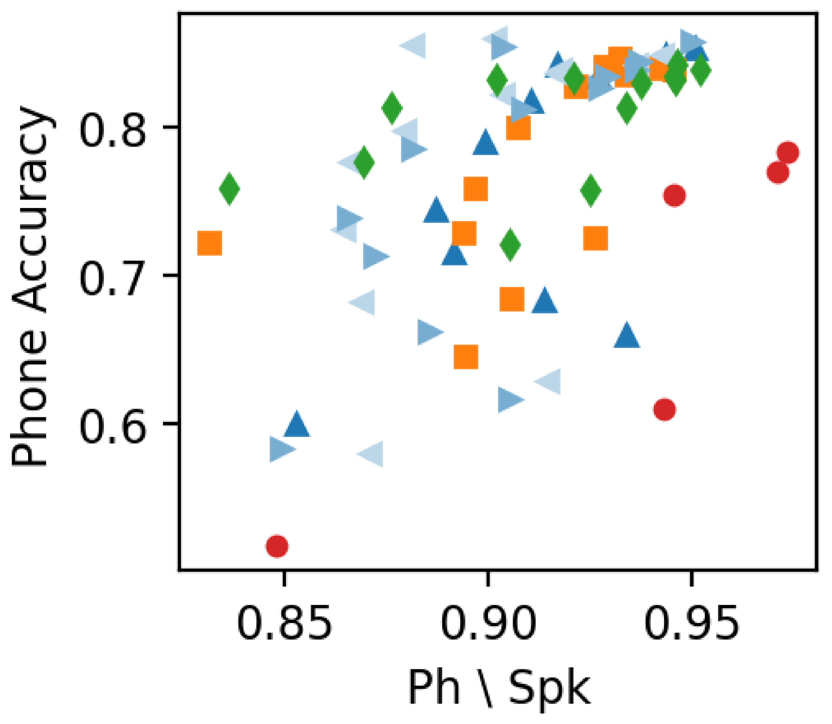


Orthogonality measured with CRV correlates with phone classification accuracy (ρ = 0.54)

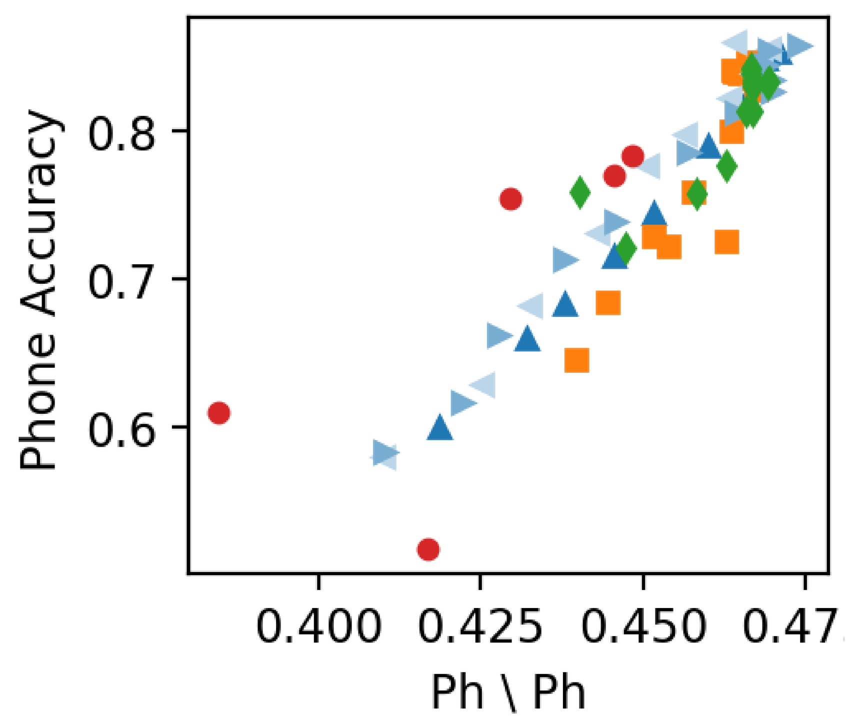
Phone classification



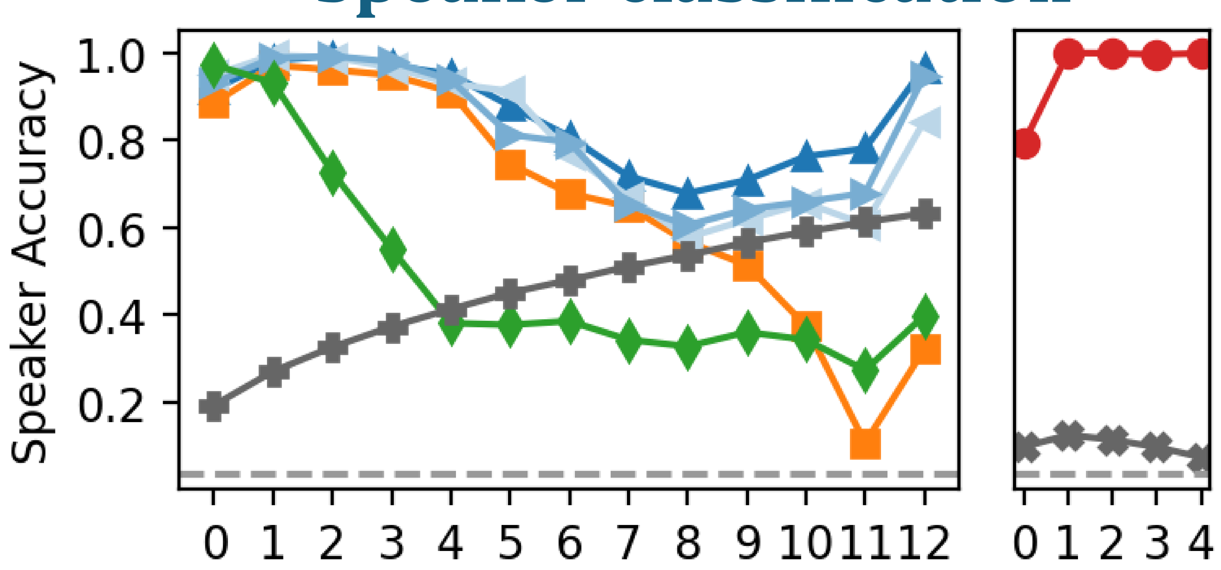
Isotropy of phone centroids showed a stronger correlation (ρ = 0.87)



The relationship between orthogonality or isotropy and speaker classification accuracy was less clear.



Speaker classification



For the untrained transformer model: both orthogonality and speaker accuracy increases across layers.

Isotropy of the frame representation space did not show consistent correlation with phone accuracy.

Across the models, layer-wise trend for speaker information shows far greater variation than phonetic information.

Cumulative residual Variance

Given **X** (speaker centroids), **Y** (phone centroids):

$X_0 = X, Y_0 = Y, v_i =$ the i-th principal direction of X.

For i = 0 to n,

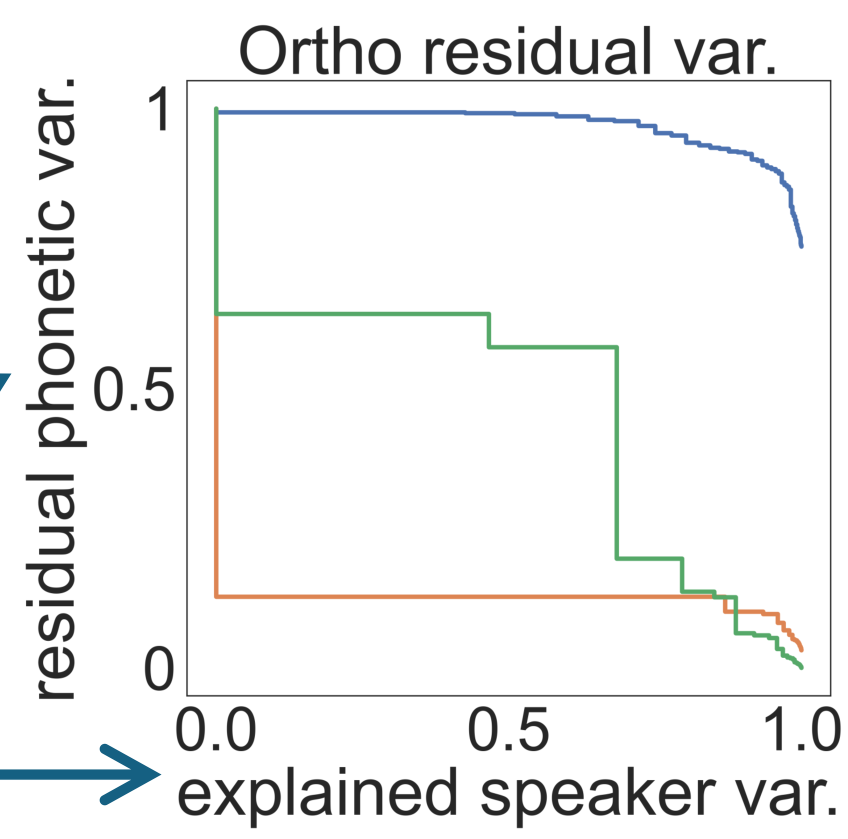
1. Project Y_i to the orthogonal complement of v_i

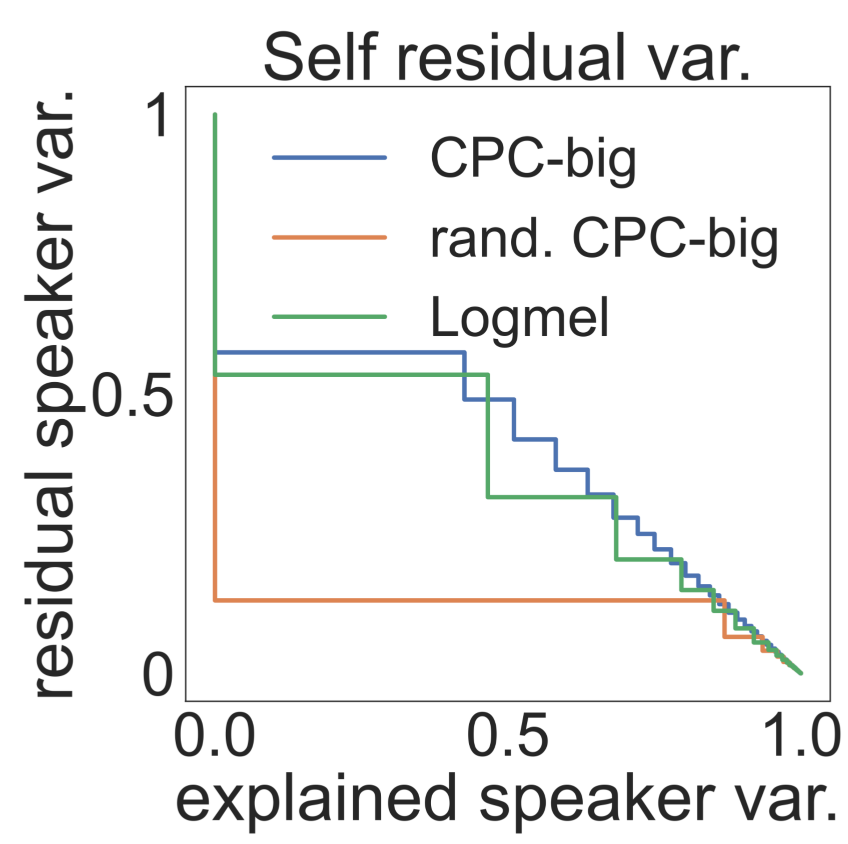
$$Y_{i+1} = Y_i - (Y_i v_i) v_i^T$$

2. Measure the variance remaining in Y_{i+1}

3. Compute the variance explained in X_{i+1} $\left(\sum_{j=1}^{i} \lambda_j\right)$

Plot





The area under the curve gives the residual phonetic variance w.r.t. speaker, or ph\spk.

AUC –> spk\spk.