Self-supervised predictive coding models encode speaker and phonetic information in orthogonal subspaces

Oli Danyi Liu, Hao Tang, Sharon Goldwater

NLP UKRI CENTRE FOR DOCTORAL TRAINING

Self-supervised predictive coding models are trained to predict a future frame given past contexts. A range of work have shown that they encode both linguistic (acoustic, word-level context ...) and non-linguistic (speaker, gender ...) information.

Open question: how are different types of information distributed across the 512 dimensions of the representation space?

Hypothesis: they are encoded orthogonally.

Specifically, we tested two models, contrastive and autoregressive predictive coding (CPC, APC):

- Architecture: both composed of LSTM layers
- Training data: 6000 (CPC) and 350 (APC) hours of speech
- Output: extract a 512-dimensional vector for each frame (1 frame per 10 ms)
- Application: used as the input to ASR models, outperform acoustic features like MFCC

To evaluate this hypothesis, we ask:

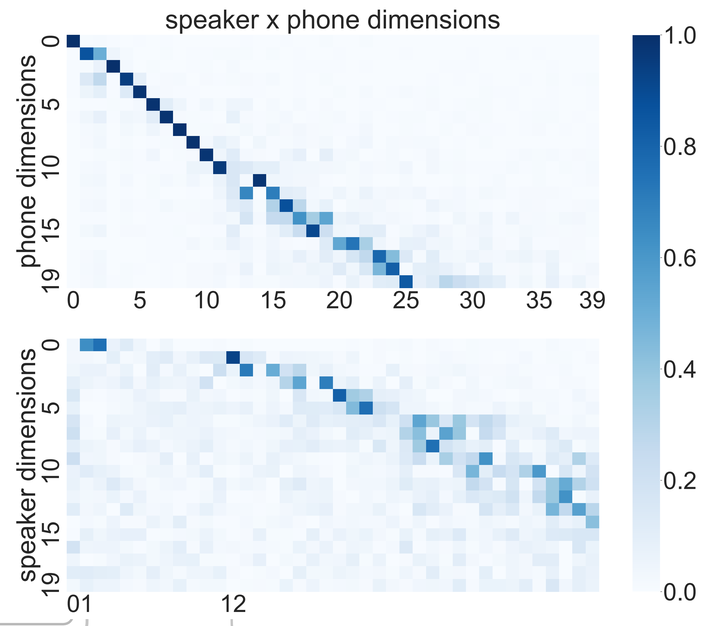
- Which dimensions in the representation space distinguish between different phones?
- Which dimensions capture speaker characteristics?
- Are they orthogonal?

We represent each speaker, each phone, each speaker x phone combination with a vector, aggregated from corresponding frame-level representations.
Our dataset for analysis contains 40 speakers, with 8-minute of speech per speaker.
Each frame is labeled with one of 39 phone labels.

Concatenating these vectors gives us:

- A speaker matrix (shape: 40 x 512)
- A phone matrix (39 x 512)
- A speaker x phone matrix (1560 x 512)

We then apply PCA to these three matrices and compare the orientation of their principal directions.
The figure below shows the similarity (absolute value of dot product) between the principal dimensions.



Amongst the top 20 speaker dimensions, their similarity with the most aligned phone dimension has mean of 0.13, variance of 0.002, and max of 0.26.
From the similarity plot, we can see that

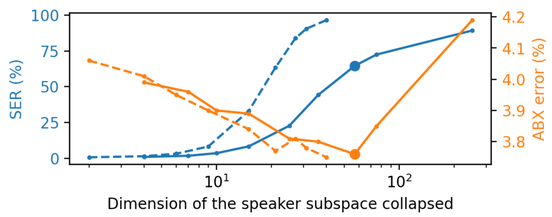
- The top 12 speaker x phone dimensions are aligned to *either* a phone *or* a speaker dimension (except for dimensions 1 and 2, although ...)
- Most of the top speaker x phone dimensions encode phone information

These are consistent with visualizations of individual dimension. Takeaways:

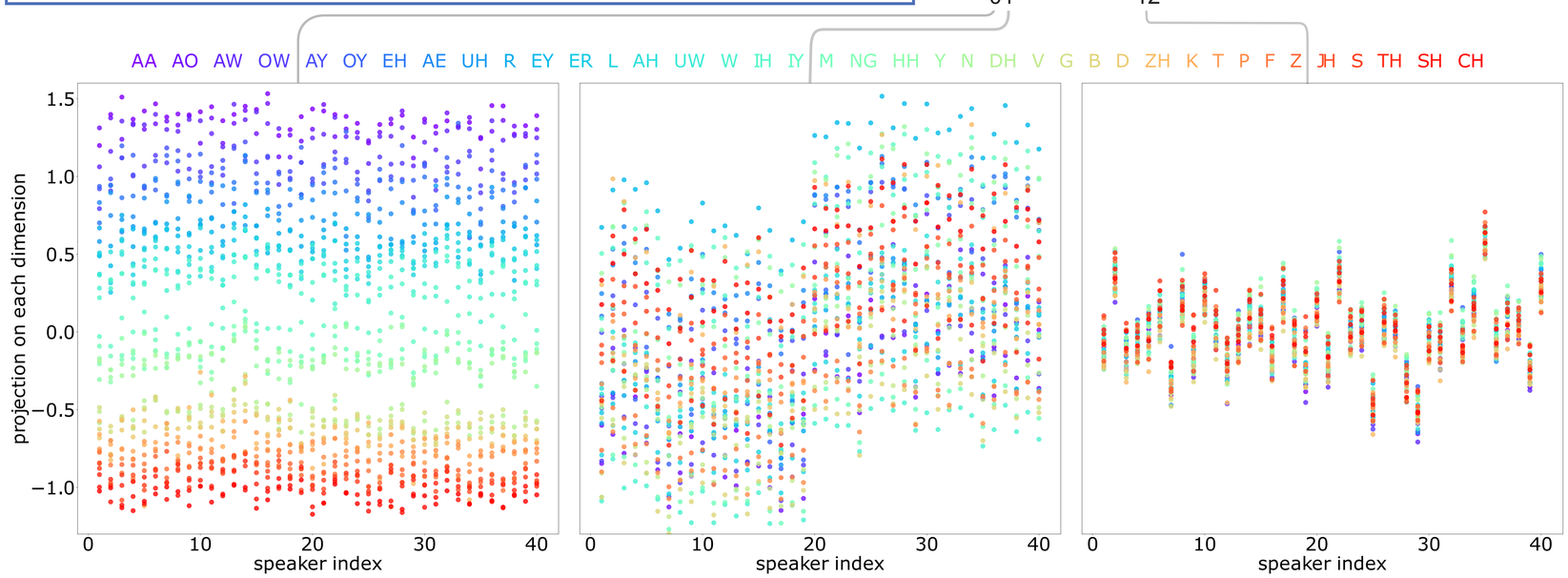
- Speaker and phone dimensions are largely orthogonal
- The representations use more dimensions to discriminate phones

Based on this property, we can perform speaker normalization by collapsing the learnt speaker subspace, no transcriptions required!

- Evaluation metrics:
  - Machine ABX (phone discrimination test): asks whether the representation of a triphone X is more similar to that of triphone token A than to B, where A and X are tokens of the same type, e.g. 'apa', and B is of a different type, e.g. 'aba'.
  - Probing classifiers: use a linear classifier to predict speaker labels based on a single frame representation. Train on a random half, use the other half for testing.
- Results:
  - For CPC, our method reduces probing accuracy from 99.5% to 4.45%, i.e. effectively eliminating speaker information, while reducing ABX error rate by 8.8%, i.e. improving phone discrimination. We observe a similar pattern for APC.
  - The method also works if we collapse the speaker subspace learnt on a different set of speakers.



Speaker error rate (SER, blue) and ABX error (orange) on our analysis dataset after collapsing the speaker subspace learnt from the same dataset (dashed) and a different dataset of 251 speakers (solid).



These figures visualize the projection on the principal dimensions 0, 1, 12 of the speaker x phone matrix. Speakers 1-19 are male and speakers 20-40 are female.