Aalto University
School of Electrical
Engineering

Interpretable Latent Space Using Space-Filling Curves for Phonetic Analysis in Voice Conversion

Mohammad Hassan Vali Tom Bäckström

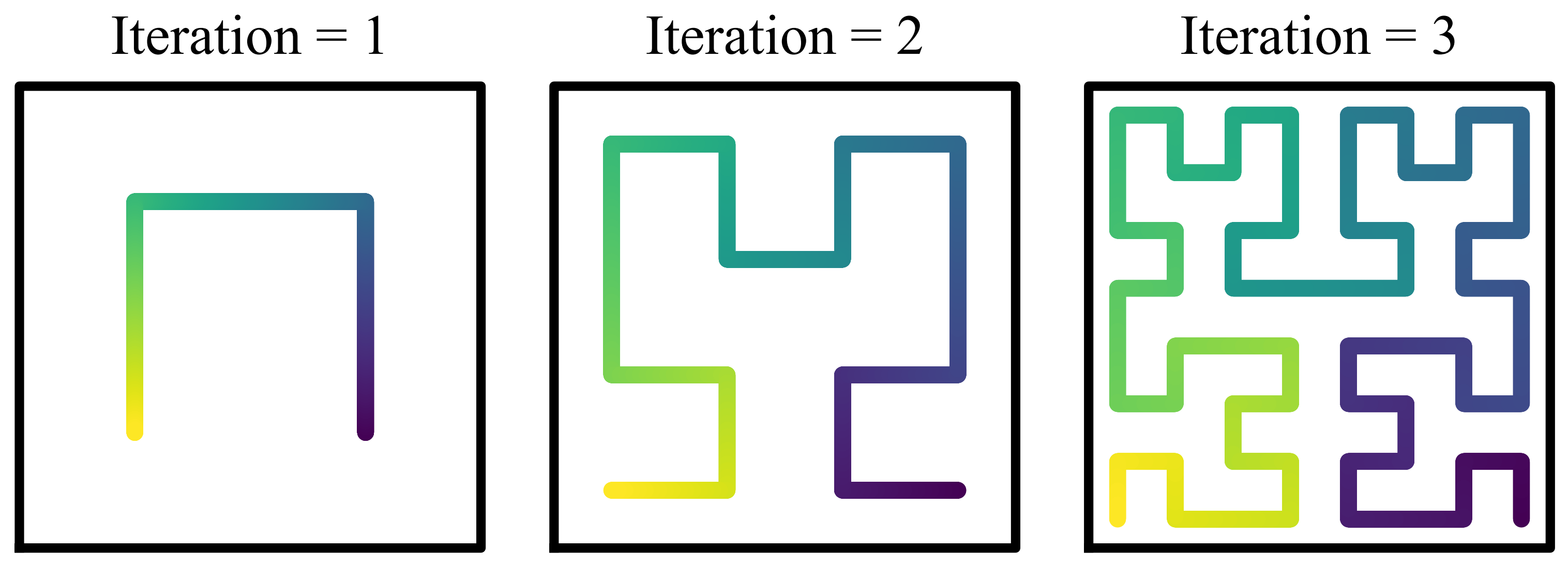
Department of Information and Communications Engineering, Aalto University, Finland

1. Introduction

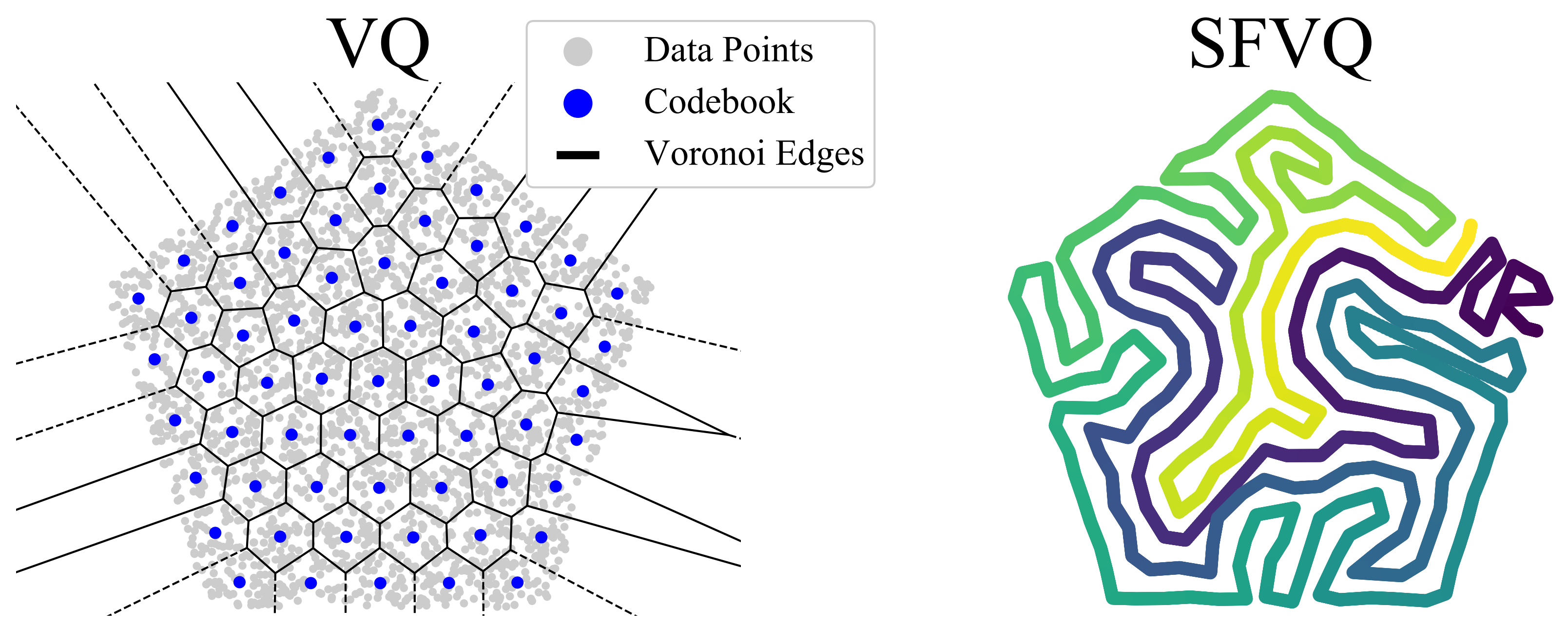
- **Central Problem:** Latent space of a deep neural network serves as a black-box, i.e. it is not interpretable.
- **Interpretable:** What info do latent vectors represent?
- **Literature:** Supervised methods learn a latent space with isolated subgroups, each representing a specific data label.
- **Disadvantages of Supervised Methods:**
  ✗ Require human labeling.
  ✗ Prevent learnt latent space to capture some inherent structures in the data.
- **Our Objective:** Explore the underlying structure in the latent space using our proposed unsupervised **Space-Filling Vector Quantizer (SFVQ)** method.

2. Proposed SFVQ Method

- **Space-Filling Curve:** A piece-wise continuous line which gets bent until it completely fills a multi-dimensional space, like this Hilbert curve:



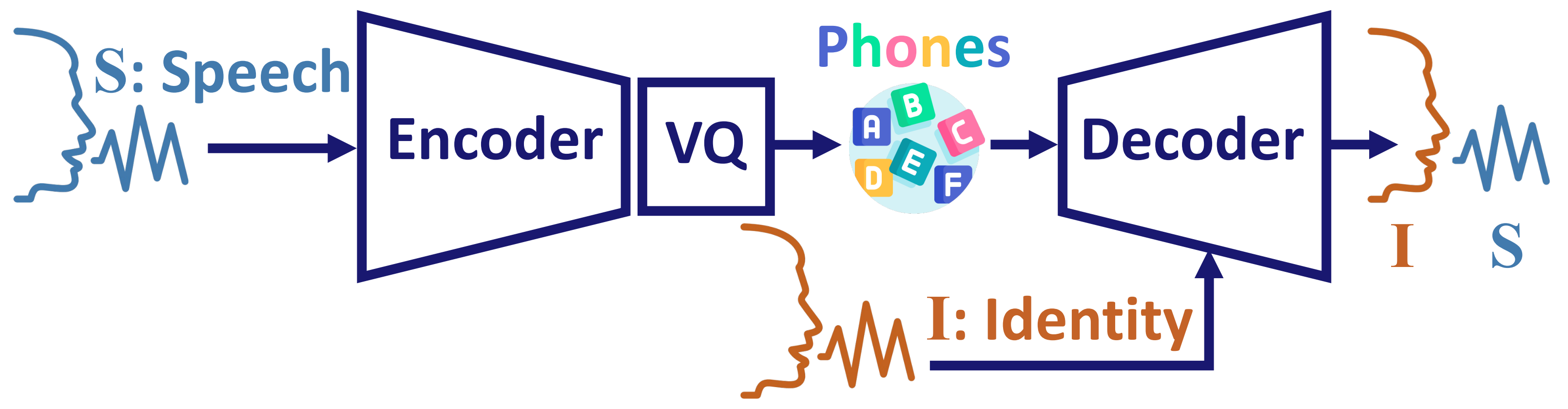
- **Vector Quantization (VQ):** A data compression technique (similar to k-means) that maps data points to codebook vectors (cluster centers).
- **Proposed Space-Filling Vector Quantizer (SFVQ):** Models a D-dimensional data distribution by a piece-wise continuous linear curve whose corner points are vector quantization codebook vectors.



- **SFVQ Advantages Over VQ:**
  ✓ *Structured codebook:* subsequent codebook vectors refer to similar contents.
  ✓ *Variable bitrate:* by selecting arbitrary equally-spaced number of points on the line as codebook vectors.
  ✓ *Higher accuracy*: possibility to decode on the lines.

3. Experimental Scenario and Setup

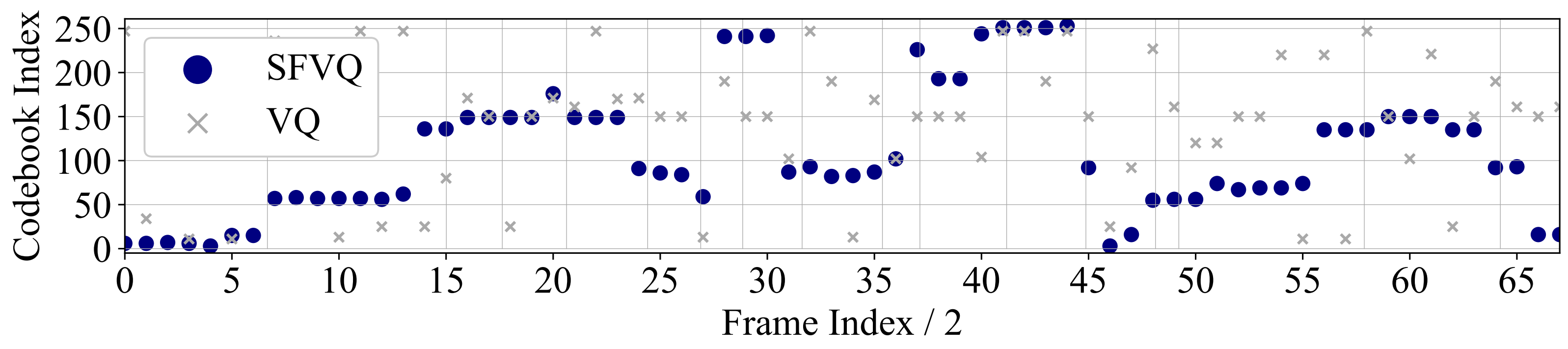
- **Framework**: Voice conversion task based on vector quantized variational autoencoder (VQ-VAE) network:



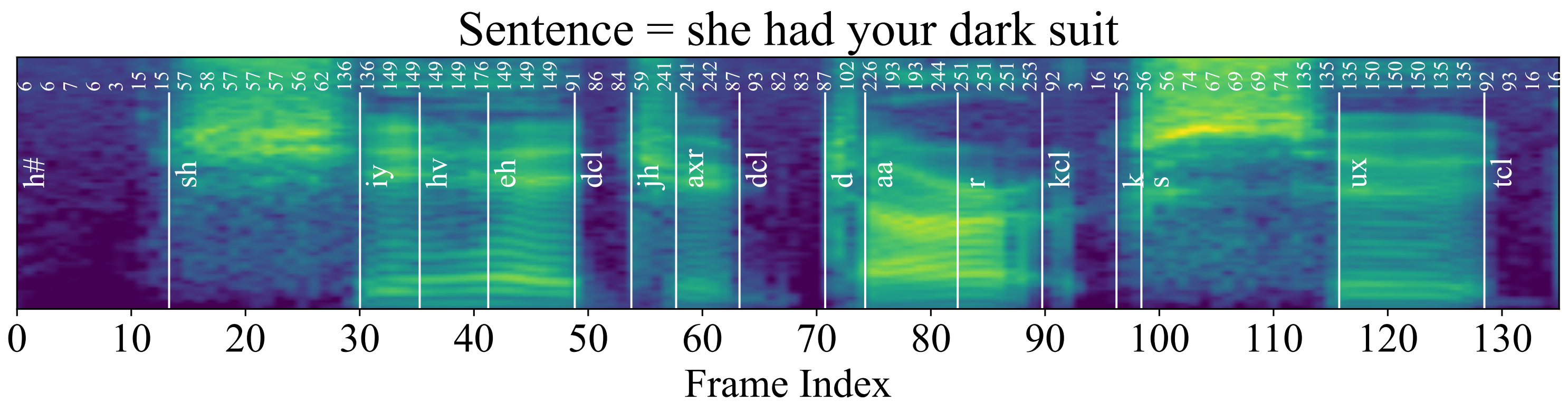
- **Vector Quantization Role**: Codebook vectors capture only phonetic content of the input speech signal.
- **Our Objective**: Explore phonetic structure in the latent space using our SFVQ, i.e. replacing VQ with SFVQ.
- **Train Data**: *ZeroSpeech* 2019 *Challenge* English dataset (15 hours of speech from 102 speakers).
- **Test Data**: *TIMIT* (phone-wise labeled dataset).
- **SFVQ Bitrate**: 8 bits = 256 codebook vectors (corner points).

4. Results

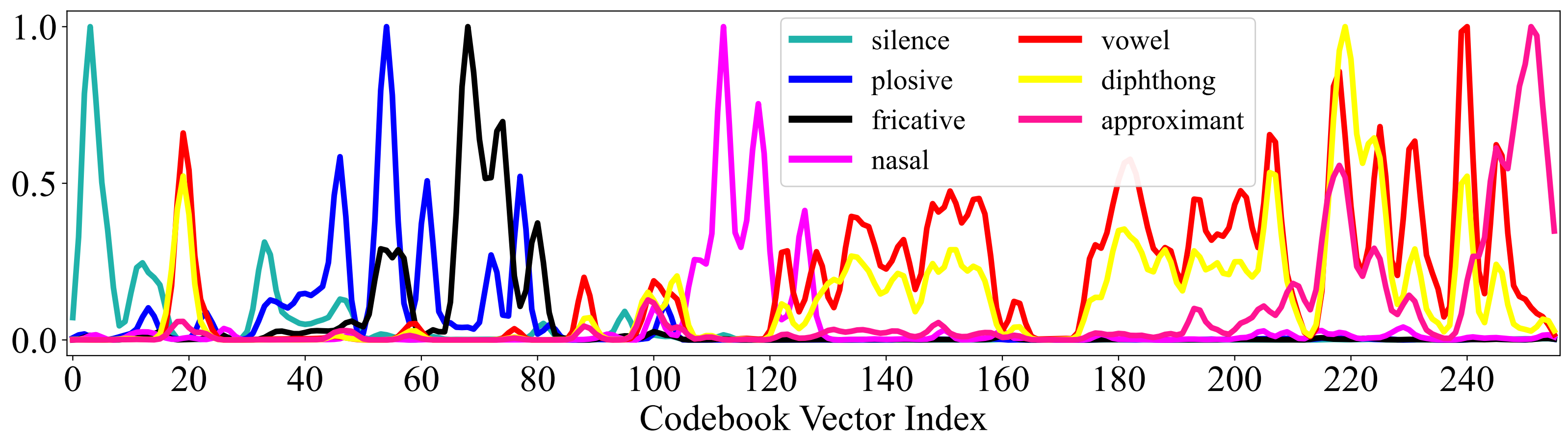
- **Examination:** Extract codebook indices for speech phones.
- **Expectation:** SFVQ maps similar phones near each other.
- VQ versus SFVQ codebook indices for a sentence:



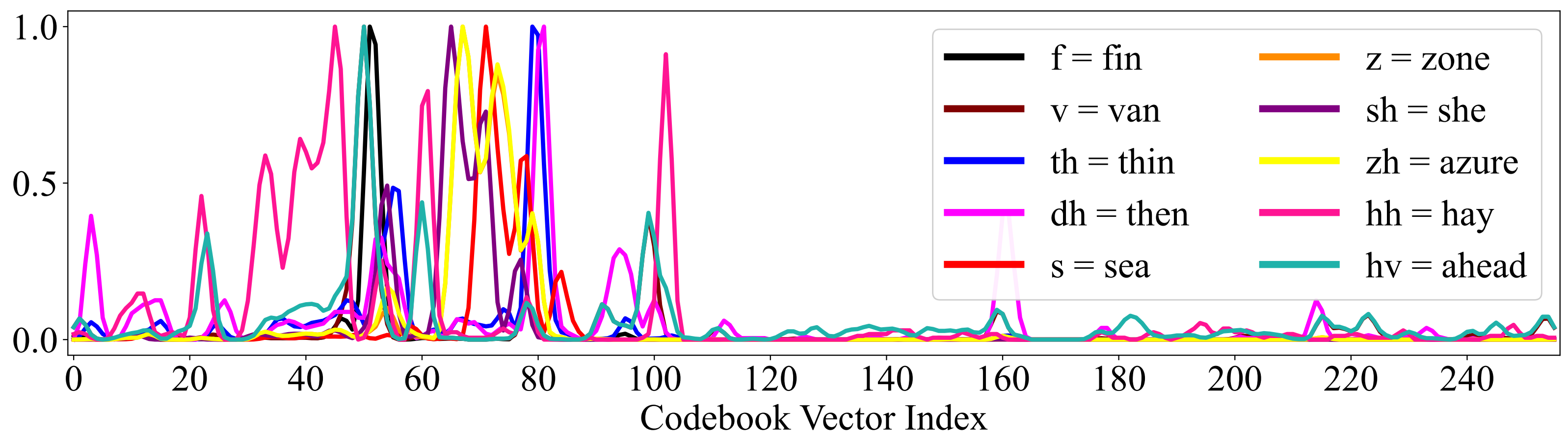
Sentence = she had your dark suit



- Histogram of codebook indices for phonetic groups:



- Histogram of codebook indices for fricative phones:



➽ **Conclusion**: Interpretable latent space by SFVQ.