S1. Supplementary Material

S1.1. Details of the FHVAE model

We follow the terminology used in [22] to describe the details of the FHVAE model. Let $\mathcal{D} = \{\boldsymbol{X^i}\}_{i=1}^{M}$ denote a speech dataset with M sequences. The i-th sequence $\boldsymbol{X^i}$ contains N^i speech segments $\{\boldsymbol{x^{(i,n)}}\}_{n=1}^{N^i}$, where $\boldsymbol{x^{(i,n)}}$ is a segment of a fixed number of frames. The FHVAE model formulates the generation process of a sequence $\boldsymbol{X}$ as[11] [22],

1. A vector $\boldsymbol{\mu_2}$ is drawn from a prior distribution $p_\theta(\boldsymbol{\mu_2}) = \mathcal{N}(\boldsymbol{0}, \sigma^2_{\boldsymbol{\mu_2}}\boldsymbol{I})$;
2. Latent segment variables $\boldsymbol{z_1^n}$ and latent sequence variables $\boldsymbol{z_2^n}$ are drawn from $p_\theta(\boldsymbol{z_1^n}) = \mathcal{N}(\boldsymbol{0}, \sigma^2_{\boldsymbol{z_1}}\boldsymbol{I})$ and $p_\theta(\boldsymbol{z_2^n}|\boldsymbol{\mu_2}) = \mathcal{N}(\boldsymbol{\mu_2}, \sigma^2_{\boldsymbol{z_2}}\boldsymbol{I})$;
3. Speech segment $\boldsymbol{x^n}$ is drawn from

$$p_\theta(\boldsymbol{x^n}|\boldsymbol{z_1^n}, \boldsymbol{z_2^n}) = \mathcal{N}(f_{\boldsymbol{\mu_x}}(\boldsymbol{z_1^n}, \boldsymbol{z_2^n}), diag(f_{\boldsymbol{\sigma_x^2}}(\boldsymbol{z_1^n}, \boldsymbol{z_2^n})). \tag{S.1}$$

Here $\mathcal{N}$ denotes the standard normal distribution, $f_{\boldsymbol{\mu_x}}(\cdot,\cdot)$ and $f_{\boldsymbol{\sigma_x^2}}(\cdot,\cdot)$ are parameterized by two DNNs. Based on Equation (S.1), the joint probability for generating $\boldsymbol{X}$ is formulated as (same as Equation (2)),

$$p_\theta(\boldsymbol{\mu_2}) \prod_{n=1}^{N} p_\theta(\boldsymbol{z_1^n}) p_\theta(\boldsymbol{z_2^n}|\boldsymbol{\mu_2}) p_\theta(\boldsymbol{x^n}|\boldsymbol{z_1^n}, \boldsymbol{z_2^n}). \tag{S.2}$$

The FHVAE introduces an inference model to approximate the true posterior as follows (same as Equation (3)),

$$p_\phi(\boldsymbol{\mu_2}) \prod_{n=1}^{N} p_\phi(\boldsymbol{z_2^n}|\boldsymbol{x^n}) p_\phi(\boldsymbol{z_1^n}|\boldsymbol{x^n}, \boldsymbol{z_2^n}). \tag{S.3}$$

Here $p_\phi(\boldsymbol{\mu_2}), p_\phi(\boldsymbol{z_2^n}|\boldsymbol{x^n})$ and $p_\phi(\boldsymbol{z_1^n}|\boldsymbol{x^n}, \boldsymbol{z_2^n})$ are all diagonal Gaussian distributions. The mean and variance values of $p_\phi(\boldsymbol{z_2^n}|\boldsymbol{x^n})$ and $p_\phi(\boldsymbol{z_1^n}|\boldsymbol{x^n}, \boldsymbol{z_2^n})$ are parameterized by DNNs.

[11] For simplicity, the superscript i in $\boldsymbol{X^i}$ and subsequent equations is omitted. This does not cause confusion.

The FHVAE is trained to optimise the *discriminative segmental variational lower bound* $\mathcal{L}(\theta, \phi; \boldsymbol{x}^{(i,n)})$ [22], which is defined as,

$$\begin{aligned}
&\mathbb{E}_{q_\phi(\boldsymbol{z}_1^{(i,n)}, \boldsymbol{z}_2^{(i,n)}|\boldsymbol{x}^{(i,n)})}[\log p_\theta(\boldsymbol{x}^{(i,n)}|\boldsymbol{z}_1^{(i,n)}, \boldsymbol{z}_2^{(i,n)})] \\
&- \mathbb{E}_{q_\phi(\boldsymbol{z}_2^{(i,n)}|\boldsymbol{x}^{(i,n)})}[\mathrm{KL}(q_\phi(\boldsymbol{z}_1^{(i,n)}|\boldsymbol{x}^{(i,n)}, \boldsymbol{z}_2^{(i,n)})||p_\theta(\boldsymbol{z}_1^{(i,n)}))] \\
&- \mathrm{KL}(q_\phi(\boldsymbol{z}_2^{(i,n)}|\boldsymbol{x}^{(i,n)})||p_\theta(\boldsymbol{z}_2^{(i,n)}|\tilde{\boldsymbol{\mu}}_2^{i})) \\
&+ \frac{1}{N^i} \log p_\theta(\tilde{\boldsymbol{\mu}}_2^{i}) + \alpha \log p(i|\boldsymbol{z}_2^{(i,n)}),
\end{aligned} \tag{S.4}$$

where $\tilde{\boldsymbol{\mu}}_2^{i}$ denotes the posterior mean of $\boldsymbol{\mu}_2$ for the i-th sequence and α denotes the discriminative weight. The discriminative objective $\log p(i|\boldsymbol{z}_2^{(i,n)})$ is formulated as,

$$\log p(i|\boldsymbol{z}_2^{(i,n)}) := \log p_\theta(\boldsymbol{z}_2^{(i,n)}|\tilde{\boldsymbol{\mu}}_2^{i}) - \log \sum_{j=1}^{M} p_\theta(\boldsymbol{z}_2^{(j,n)}|\tilde{\boldsymbol{\mu}}_2^{j}). \tag{S.5}$$

After FHVAE training, $\boldsymbol{z}_1$ representation is extracted as the desired speaker-invariant representation of speech.