A PROOFS

Theorem A.1 (Theorem 3.1)**.** *Let* $\boldsymbol{\mu}_v^{(-ui)} = \frac{1}{N_v}\sum_{k=1}^{N_v} \boldsymbol{s}_{vk}$ *if* $u \neq v$; *and* $\boldsymbol{\mu}_u^{(-ui)} = \frac{1}{N_u - 1}\sum_{j \neq i} \boldsymbol{s}_{uj}$*. Then,*

$$\mathcal{I}(\boldsymbol{u};\boldsymbol{s}) \geq \mathbb{E}\Big[\frac{1}{N}\sum_{u=1}^{M}\sum_{i=1}^{N_u}\Big[- \|\boldsymbol{s}_{ui} - \boldsymbol{\mu}_u^{(-ui)}\|^2 - \frac{e^{-1}}{N}\sum_{v=1}^{M}[N_v \exp(-\|\boldsymbol{s}_{ui} - \boldsymbol{\mu}_v^{(-ui)}\|^2)]\Big]\Big]. \tag{12}$$

Proof of Theorem 3.1. By the condition in the Theorem, we have the given sample pairs $\{(u, \boldsymbol{s}_{ui})\}_{1\leq u \leq M, 1 \leq i \leq N_u}$. Not that each pair of speaker identity and style embedding, $(u, \boldsymbol{s}_{ui})$, can be regarded as a sample from the joint distribution $p(\boldsymbol{u}, \boldsymbol{s})$. To clearify the proof, we change the notation of random variables $\boldsymbol{u}$ and $\boldsymbol{s}$ to $\boldsymbol{U}$ and $\boldsymbol{S}$, which are distinct to samples $\{(u, \boldsymbol{s}_{ui})\} \sim p(\boldsymbol{U}, \boldsymbol{S})$.

For a sample pair $(u, \boldsymbol{s}_{ui})$, by the NWJ lower bound, we have

$$\begin{aligned}\mathcal{I}(\boldsymbol{U};\boldsymbol{S}) \geq & \mathbb{E}_{p(\boldsymbol{U},\boldsymbol{S})}[f(\boldsymbol{U},\boldsymbol{S})] - e^{-1}\mathbb{E}_{p(\boldsymbol{U})p(\boldsymbol{S})}[e^{f(\boldsymbol{U},\boldsymbol{S})}] \\ = & \mathbb{E}_{p(\boldsymbol{S})}\Big[\mathbb{E}_{p(\boldsymbol{U}|\boldsymbol{S})}[f(\boldsymbol{U},\boldsymbol{S})] - e^{-1}\mathbb{E}_{p(\boldsymbol{U})}[e^{f(\boldsymbol{U},\boldsymbol{S})}]\Big] \\ = & \mathbb{E}_{\boldsymbol{s}_{ui}\sim p(\boldsymbol{S})}\Big[\mathbb{E}_{p(\boldsymbol{U}|\boldsymbol{S}=\boldsymbol{s}_{ui})}[f(\boldsymbol{U},\boldsymbol{S}=\boldsymbol{s}_{ui})] - e^{-1}\mathbb{E}_{p(\boldsymbol{U})}[e^{f(\boldsymbol{U},\boldsymbol{S}=\boldsymbol{s}_{ui})}]\Big],\end{aligned} \tag{13}$$

with a score function $f(\boldsymbol{U}, \boldsymbol{S})$. Given $\boldsymbol{S} = \boldsymbol{s}_{ui}$, $\mathbb{E}_{p(\boldsymbol{U}|\boldsymbol{S}=\boldsymbol{s}_{ui})}[f(\boldsymbol{U},\boldsymbol{S}=\boldsymbol{s}_{ui})]$ has an unbiased estimation $f(\boldsymbol{U} = u, \boldsymbol{S} = \boldsymbol{s}_{ui})$; $\mathbb{E}_{p(\boldsymbol{U})}[e^{f(\boldsymbol{U},\boldsymbol{S}=\boldsymbol{s}_{ui})}]$ has an unbiased estimation by taking average of all possible values $v \sim p(\boldsymbol{U})$ in samples $\{(u, \boldsymbol{s}_{ui})\}$,

$$\mathbb{E}_{p(\boldsymbol{U})}[e^{f(\boldsymbol{U},\boldsymbol{S}=\boldsymbol{s}_{ui})}] = \mathbb{E}\Big[\sum_{v=1}^{M}\frac{N_v}{N}e^{f(\boldsymbol{U}=v,\boldsymbol{S}=\boldsymbol{s}_{ui})}\Big]. \tag{14}$$

With the two estimations, (13) becomes

$$\mathcal{I}(\boldsymbol{U};\boldsymbol{S}) \geq \mathbb{E}\Big[f(u,\boldsymbol{s}_{ui}) - e^{-1}\sum_{v=1}^{M}\frac{N_v}{N}e^{f(v,\boldsymbol{s}_{ui})}\Big]. \tag{15}$$

Specifically, we select score function $f(\boldsymbol{U} = v, \boldsymbol{S} = \boldsymbol{s}) = -\|\boldsymbol{s} - \boldsymbol{\mu}_v^{(-ui)}\|^2$, then (15) becomes

$$\mathcal{I}(\boldsymbol{U};\boldsymbol{S}) \geq \mathbb{E}[\hat{\mathcal{I}}_{ui}] = \mathbb{E}\Big[- \|\boldsymbol{s}_{ui} - \boldsymbol{\mu}_u^{-(ui)}\|^2 - e^{-1}\sum_{v=1}^{M}\frac{N_v}{N}e^{-\|\boldsymbol{s}_{ui} - \boldsymbol{\mu}_v^{(-ui)}\|^2}\Big]. \tag{16}$$

Since the selection of index ui is arbitrary, we take an average on all $\hat{\mathcal{I}}_{ui}$,

$$\begin{aligned}\mathcal{I}(\boldsymbol{U};\boldsymbol{S}) \geq & \frac{1}{N}\sum_{u=1}^{M}\sum_{i=1}^{N_u}\mathbb{E}_{(u,\boldsymbol{s}_{ui})\sim p(\boldsymbol{U},\boldsymbol{S})}[\hat{\mathcal{I}}_{ui}] = \mathbb{E}[\frac{1}{N}\sum_{u=1}^{M}\sum_{i=1}^{N_u}\hat{\mathcal{I}}_{ui}] \\ = & \mathbb{E}\Big[\frac{1}{N}\sum_{u=1}^{M}\sum_{i=1}^{N_u}\Big[- \|\boldsymbol{s}_{ui} - \boldsymbol{\mu}_u^{(-ui)}\|^2 - \frac{e^{-1}}{N}\sum_{v=1}^{M}[N_v \exp(-\|\boldsymbol{s}_{ui} - \boldsymbol{\mu}_v^{(-ui)}\|^2)]\Big]\Big],\end{aligned} \tag{17}$$

where the right-hand side of equation (7) is derived. □

Theorem A.2 (Theorem 3.2)**.** *Assume that given* $\boldsymbol{s} = \boldsymbol{s}_u$*, samples* $\{(\boldsymbol{x}_{ui}, \boldsymbol{c}_{ui})\}_{i=1}^{N_u}$ *are observed. With a variational distribution* $q_\phi(\boldsymbol{x}|\boldsymbol{s}, \boldsymbol{c})$*, we have* $\mathcal{I}(\boldsymbol{x}; \boldsymbol{c}|\boldsymbol{s}) \geq \mathbb{E}[\hat{\mathcal{I}}]$*, where*

$$\hat{\mathcal{I}} = \frac{1}{N}\sum_{u=1}^{M}\sum_{i=1}^{N_u}\Big[\log q_\phi(\boldsymbol{x}_{ui}|\boldsymbol{c}_{ui}, \boldsymbol{s}_u) - \log(\frac{1}{N_u}\sum_{j=1}^{N_u} q_\phi(\boldsymbol{x}_{uj}|\boldsymbol{c}_{ui}, \boldsymbol{s}_u))\Big]. \tag{18}$$

Proof of Theorem 3.2 . Given $\boldsymbol{s} = \boldsymbol{s}_u$, we observe sample pair $\{\boldsymbol{x}_{ui}, \boldsymbol{c}_{ui}\}_{i=1}^{N_u}$. By the InfoNCE lower bound (Oord et al., 2018), with a score function f, we have

$$\mathcal{I}(\boldsymbol{x};\boldsymbol{c}|\boldsymbol{s} = \boldsymbol{s}_u) \geq \mathbb{E}\Big[\frac{1}{N_u}\sum_{i=1}^{N_u}\Big[f(\boldsymbol{x}_{ui}, \boldsymbol{c}_{ui}) - \log\left(\frac{1}{N_u}\sum_{j=1}^{N_u}e^{f(\boldsymbol{x}_{uj},\boldsymbol{c}_{ui})}\right)\Big]\Big]. \tag{19}$$

We select $f(\boldsymbol{x},\boldsymbol{c}) = \log q_\phi(\boldsymbol{x}|\boldsymbol{c}, \boldsymbol{s}=\boldsymbol{s}_u)$, then

$$\mathcal{I}(\boldsymbol{x};\boldsymbol{c}|\boldsymbol{s}=\boldsymbol{s}_u) \geq \mathbb{E}\Big[\frac{1}{N_u}\sum_{i=1}^{N_u}\Big[\log q_\phi(\boldsymbol{x}_{ui}|\boldsymbol{c}_{ui},\boldsymbol{s}_u) - \log\left(\frac{1}{N_u}\sum_{j=1}^{N_u} q_\phi(\boldsymbol{x}_{uj}|\boldsymbol{c}_{ui},\boldsymbol{s}_u)\right)\Big]\Big]. \tag{20}$$

Taking expectation of $\boldsymbol{s}$ on both sides, we derive

$$\mathcal{I}(\boldsymbol{x};\boldsymbol{c}|\boldsymbol{s}) \geq \mathbb{E}\Big[\frac{1}{N}\sum_{u=1}^{M}\sum_{i=1}^{N_u}\Big[\log q_\phi(\boldsymbol{x}_{ui}|\boldsymbol{c}_{ui},\boldsymbol{s}_u) - \log\left(\frac{1}{N_u}\sum_{j=1}^{N_u} q_\phi(\boldsymbol{x}_{uj}|\boldsymbol{c}_{ui},\boldsymbol{s}_u)\right)\Big]\Big]. \tag{21}$$

□

Theorem A.3 (Theorem 3.3). *If $p(\boldsymbol{s}|\boldsymbol{c})$ provides the conditional distribution between variables $\boldsymbol{s}$ and $\boldsymbol{c}$, then*

$$\mathcal{I}(\boldsymbol{s};\boldsymbol{c}) \leq \mathbb{E}\Big[\frac{1}{N}\sum_{u=1}^{M}\sum_{i=1}^{N_u}\Big[\log p(\boldsymbol{s}_{ui}|\boldsymbol{c}_{ui}) - \frac{1}{N}\sum_{v=1}^{M}\sum_{j=1}^{N_v}\log p(\boldsymbol{s}_{ui}|\boldsymbol{c}_{vj})\Big]\Big]. \tag{22}$$

Proof of Theorem 3.3. By the upper bound in Cheng *et al.* (Cheng et al., 2020b), we have

$$\mathcal{I}(\boldsymbol{s};\boldsymbol{c}) \leq \mathbb{E}_{p(\boldsymbol{s},\boldsymbol{c})}[\log p(\boldsymbol{s}|\boldsymbol{c})] - \mathbb{E}_{p(\boldsymbol{s})p(\boldsymbol{c})}[\log p(\boldsymbol{s}|\boldsymbol{c})]. \tag{23}$$

With embedding samples $\{\boldsymbol{s}_{ui},\boldsymbol{c}_{ui}\}_{1\leq u\leq M, 1\leq i\leq N_u}$, the right-hand side of (23) can be estimated by

$$\mathcal{I}(\boldsymbol{s};\boldsymbol{c}) \leq \mathbb{E}\Big[\frac{1}{N}\sum_{u=1}^{M}\sum_{i=1}^{N_u}\Big[\log p(\boldsymbol{s}_{ui}|\boldsymbol{c}_{ui}) - \frac{1}{N}\sum_{v=1}^{M}\sum_{j=1}^{N_v}\log p(\boldsymbol{s}_{ui}|\boldsymbol{c}_{vj})\Big]\Big]. \tag{24}$$

□

Discussion on variational approximation As mentioned in Section 3.3, we approximate $p(\boldsymbol{s}|\boldsymbol{c})$ with a variational distribution $q_\theta(\boldsymbol{s}|\boldsymbol{c})$ in equation (10), since the closed form of $p(\boldsymbol{s}|\boldsymbol{c})$ is unknown. We claim that with $q_\theta(\boldsymbol{s}|\boldsymbol{c})$ as a good approximation of $p(\boldsymbol{s}|\boldsymbol{c})$, equation (10) remains a MI upper bound. We calculate the difference between $\mathcal{I}(\boldsymbol{s};\boldsymbol{c})$ and the approximated version of (23):

$$\begin{aligned}
\Delta :=& \mathcal{I}(\boldsymbol{s};\boldsymbol{c}) - [\mathbb{E}_{p(\boldsymbol{s},\boldsymbol{c})}[\log q_\theta(\boldsymbol{s}|\boldsymbol{c})] - \mathbb{E}_{p(\boldsymbol{s})p(\boldsymbol{c})}[\log q_\theta(\boldsymbol{s}|\boldsymbol{c})]] \\
=& \mathbb{E}_{p(\boldsymbol{s},\boldsymbol{c})}[\log p(\boldsymbol{s}|\boldsymbol{c}) - \log p(\boldsymbol{s})] - \mathbb{E}_{p(\boldsymbol{s},\boldsymbol{c})}[\log q_\theta(\boldsymbol{s}|\boldsymbol{c})] + \mathbb{E}_{p(\boldsymbol{s})p(\boldsymbol{c})}\Big[\log q_\theta(\boldsymbol{s}|\boldsymbol{c})\Big] \\
=& \Big[\mathbb{E}_{p(\boldsymbol{s},\boldsymbol{c})}[\log p(\boldsymbol{s}|\boldsymbol{c})] - \mathbb{E}_{p(\boldsymbol{s},\boldsymbol{c})}[\log q_\theta(\boldsymbol{s}|\boldsymbol{c})]\Big] - \Big[\mathbb{E}_{p(\boldsymbol{s})}[\log p(\boldsymbol{s})] - \mathbb{E}_{p(\boldsymbol{s})p(\boldsymbol{c})}[\log q_\theta(\boldsymbol{s}|\boldsymbol{c})]\Big] \\
=& \mathbb{E}_{p(\boldsymbol{s},\boldsymbol{c})}[\log\frac{p(\boldsymbol{s}|\boldsymbol{c})}{q_\theta(\boldsymbol{s}|\boldsymbol{c})}] - \mathbb{E}_{p(\boldsymbol{s})p(\boldsymbol{c})}[\log\frac{p(\boldsymbol{s})}{q_\theta(\boldsymbol{s}|\boldsymbol{c})}] \\
=& \mathrm{KL}(p(\boldsymbol{s}|\boldsymbol{c})\|q_\theta(\boldsymbol{s}|\boldsymbol{c})) - \mathrm{KL}(p(\boldsymbol{s})\|q_\theta(\boldsymbol{s}|\boldsymbol{c})).
\end{aligned}$$

When $q_\theta(\boldsymbol{s}|\boldsymbol{c})$ is a good approximation to $p(\boldsymbol{s}|\boldsymbol{c})$, the divergence $\mathrm{KL}(p(\boldsymbol{s}|\boldsymbol{c})\|q_\theta(\boldsymbol{s}|\boldsymbol{c}))$ can be smaller than $\mathrm{KL}(p(\boldsymbol{s})\|q_\theta(\boldsymbol{s}|\boldsymbol{c}))$. Then Δ remains negative, which indicates $[\mathbb{E}_{p(\boldsymbol{s},\boldsymbol{c})}[\log q_\theta(\boldsymbol{s}|\boldsymbol{c})] - \mathbb{E}_{p(\boldsymbol{s})p(\boldsymbol{c})}[\log q_\theta(\boldsymbol{s}|\boldsymbol{c})]]$ still be an MI upper bound.

B EXPERIMENTS

More ablation study on bottleneck design We kept the same bottleneck design as AUTOVC to have a fair comparison for the effectiveness of the proposed disentangled learning scheme. To further provide evidence of effectiveness of IDE-VC , we also conducted an ablation study in which the bottleneck is widened in Table 5. Specifically, we use set sampling rate as 4 and conduct experiments under the zero-shot setup. The results demonstrated that the bottleneck design has little impact on the disentanglement ability of the proposed model.

Table 5: Ablation study with bottleneck design for zero-shot VST. We set sampling rate as 4. Performance is measured by objective metrics.

	Distance	Verification[%]
AUTOVC	6.59	41
IDE-VC	**6.24**	**80**

More ablation study on visualization We further provide t-SNE visualization for content embedding from IDE-VC and AUTOVC in Figure 3 and Figure 4 under same hyperparameter setups. Comparing between the two t-SNE plottings, the content embeddings generated with IDE-VC are more indistinguishable for different speakers than the ones from AUTOVC, which proves that the proposed model has stronger ability to eliminate speaker-related information in content embedding.

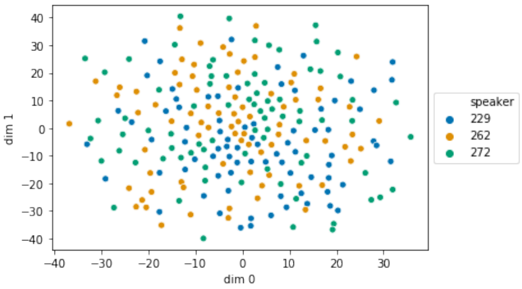


Figure 3: t-SNE visualization for content embedding from IDE-VC . The embeddings are extracted from the voice samples of 3 different speakers.

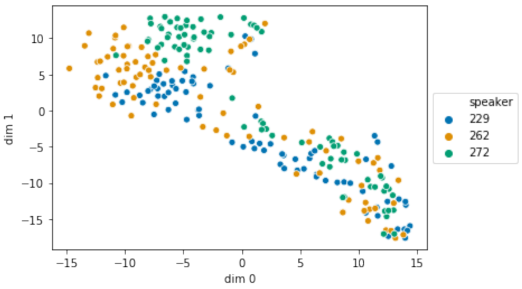


Figure 4: t-SNE visualization for content embedding from AUTOVC. The embeddings are extracted from the voice samples of 3 different speakers.

Speaker encoder pretraining Our speaker encoder is pretrained with GE2E loss on a combination of VoxCeleb1 (Nagrani et al., 2017) and Librispeech (Panayotov et al., 2015) datasets, in total of 3549 speakers.

Implementation details In our experiments, we use official implementation of AdaIN-VC[4], AUTOVC[5] and Blow[6]. Specifically, same pretrained speaker encoder is used in AUTOVC (Qian et al., 2019) and our model for fair comparison. Blow model is trained with 100 epochs and suggested hyperparameters, the training takes over 10 GPU days on Nvidia V100 in comparison with 1 GPU day on Nvidia Xp for our model. For StarGAN-VC, we use an open source implementation[7], which achieves better performance according to multiple previous works (Qian et al., 2019; Serrà et al., 2019). All above models are trained on all 109 speakers in VCTK dataset, and same splits are used for testing and validation.

For our model, we use loss on validation set to conduct grid search on hyperparameter β, and we use $\beta = 5$ in final experiment. The other hyperparameters are set as the same as in AUTOVC (Qian et al., 2019).

Sample speeches We also provide several sample conversed speeches on https://idevc.github.io/.

---

[4] https://github.com/jjery2243542/adaptive_voice_conversion

[5] https://github.com/auspicious3000/autovc

[6] https://github.com/joansj/blow

[7] https://github.com/liusongxiang/StarGAN-Voice-Conversion

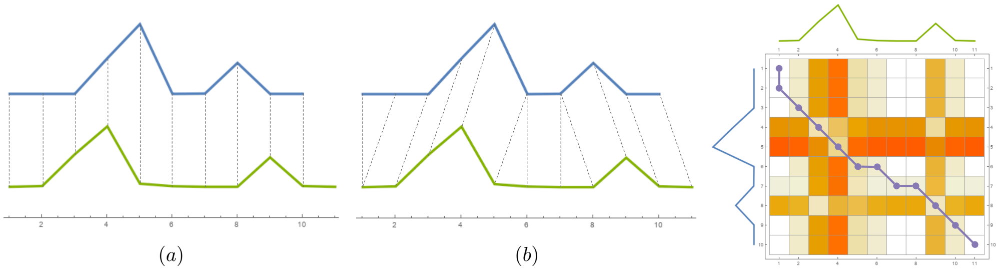


Figure 5: Left: (a) The naive Euclidean distance matching between two time series might miss the important voice patterns because of the time shift and different speakers' speaking speeds; (b) DTW automatically find the optimal matching that captures important voice pattens. Right: DTW converts the time sequence matching problem into the minimal cost path searching on the distance matrix.

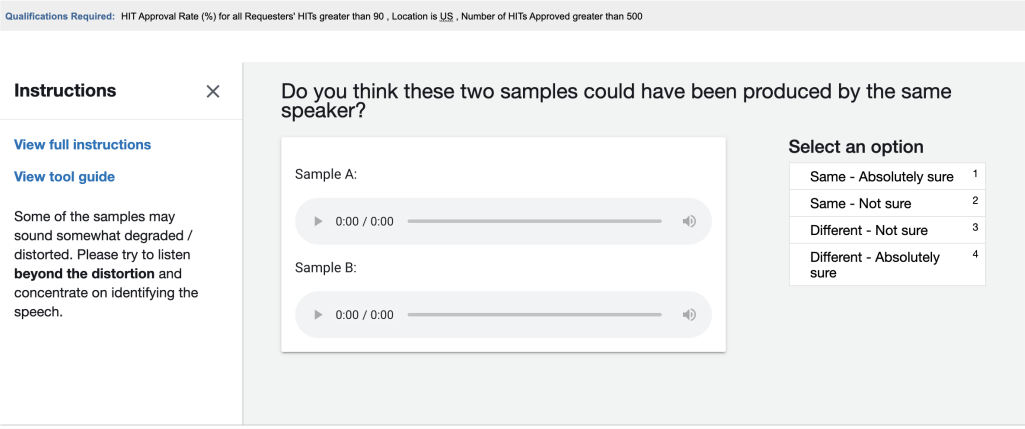
C EVALUATION DETAILS

Verification score In details, in Resemblyzer, a pre-trained speaker encoder is provided $\boldsymbol{s} = E_{sr}(\boldsymbol{x})$. The voice profile for each speaker u is first computed as $\boldsymbol{s}_u = \frac{1}{N_t}\sum_n^{N_t} E_{sr}(\boldsymbol{x}_{un})$, in which N_t represents the number of speeches each speaker has in testing set, $\boldsymbol{x}_{un}$ represents the n-th speech of speaker u in testing set. For each speech conversed from i-th speech of speaker u to j-th speech of speaker v, represented as $\hat{\boldsymbol{x}}_{u_i \to v_j}$, the speaker embedding is computed with Resemblyzer: $\hat{\boldsymbol{s}}_{u_i \to v_j} = E_{sr}(\hat{\boldsymbol{x}}_{u_i \to v_j})$. Dot product is used to compute similarity between the speaker embedding and the voice profile. If among all speakers in testing set, the speaker embedding of the conversed speech has highest similarity score with the target speaker's profile $\boldsymbol{s}_v$, we view it as a success conversion. The portion of success conversion among all conversion trials is reported as verification score.

Details in evaluation for zero-shot VST Based on setting in AdaIN-VC (Chou & Lee, 2019), subjective evaluation is performed on converted voice between male to male, male to female, female to male and female to female speakers. To reduce variance, 3 speakers are selected for each gender, thus, in total 36 pairs of speakers. The speakers of these pairs were unseen during training. Following the setting in AdaIN-VC (Chou & Lee, 2019), the converted result of each pair was transfered from our proposed model with only one source utterance and one target utterance.

Dynamic Time Wrapping When evaluating the voices generated by neural networks from latent embeddings, the mismatching problem in time alignment occurs due to the time shift and different speaking speeds in the generation. Important voice patterns may be neglected when directly calculating the Euclidean distance between the generated voice and the ground-truth. One effective solution to the sequential time-alignment problem is the Dynamic Time Wrapping (DTW) algorithm (Berndt & Clifford, 1994), which has been widely applied in speech recognition and matching (Muda et al., 2010; Chapaneri, 2012; Dhingra et al., 2013). The DTW algorithm seeks the optimal matching path $\boldsymbol{P}^* \in \mathcal{P}(T, S)$ that minimizes the sequential matching cost between two time series $\boldsymbol{x} = (x^1, x^2, \ldots, x^T)$ and $\boldsymbol{y} = (y^1, y^2, \ldots, y^S)$ (*e.g.*, the purple path in the right of Figure 5). A consecutive matching path $\boldsymbol{P} \in \mathcal{P}(T, S)$ denotes a sequence of index pairs $\boldsymbol{P} = (\boldsymbol{p}^1, \boldsymbol{p}^2, \ldots, \boldsymbol{p}^L)$, in which each pair $\boldsymbol{p}^l = (t_l, s_l)$ matches x^{t_l} and y^{s_l} following the time order (*i.e.*, $\boldsymbol{p}^1 = (1, 1)$, $\boldsymbol{p}^L = (T, S)$, $0 \leq t_{l+1} - t_l \leq 1$, and $0 \leq s_{l+1} - s_l \leq 1$). The DTW score is $\mathcal{S}_{\text{DTW}}(\boldsymbol{x}, \boldsymbol{y}) = \min_{\boldsymbol{P} \in \mathcal{P}(T,S)} \sum_{l=1}^{L} d(x^{t_l}, y^{s_l})$, where $d(\cdot, \cdot)$ is a ground distance measuring the dissimilarity of any two points in time series. The optimization needed for calculating the DTW score can be solved efficiently by dynamic programming.

Human Evaluation Following Wester *et al.* (Wester et al., 2016), we use the naturalness of the speech and the similarity of the transferred speech to target identity as subjective metrics. Figure 6 and Figure 7 shows the contents of the two human evaluation webpage layouts respectively.



Qualifications Required: HIT Approval Rate (%) for all Requesters' HITs greater than 90 , Location is US , Number of HITs Approved greater than 500

Instructions

View full instructions

View tool guide

Some of the samples may sound somewhat degraded / distorted. Please try to listen **beyond the distortion** and concentrate on identifying the speech.

Do you think these two samples could have been produced by the same speaker?

Sample A:

0:00 / 0:00

Sample B:

0:00 / 0:00

Select an option

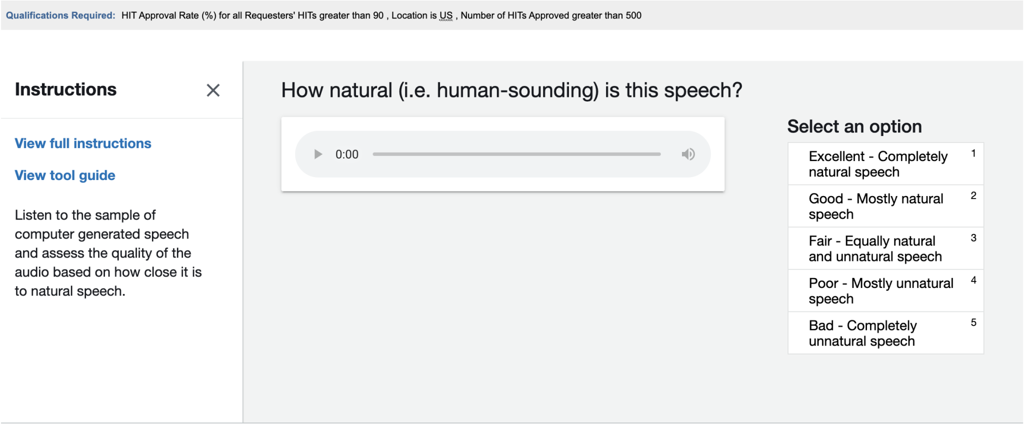
Same - Absolutely sure 1

Same - Not sure 2

Different - Not sure 3

Different - Absolutely sure 4

Figure 6: Human evaluation: similarity



Qualifications Required: HIT Approval Rate (%) for all Requesters' HITs greater than 90 , Location is US , Number of HITs Approved greater than 500

Instructions

View full instructions

View tool guide

Listen to the sample of computer generated speech and assess the quality of the audio based on how close it is to natural speech.

How natural (i.e. human-sounding) is this speech?

0:00

Select an option

Excellent - Completely natural speech 1

Good - Mostly natural speech 2

Fair - Equally natural and unnatural speech 3

Poor - Mostly unnatural speech 4

Bad - Completely unnatural speech 5

Figure 7: Human evaluation: naturalness

D DATA PROCESSING INEQUALITY

Theorem D.1. *If three variables $\boldsymbol{x} \to \boldsymbol{y} \to \boldsymbol{z}$ follow a markov chain, then $\mathcal{I}(\boldsymbol{x};\boldsymbol{y}) \geq \mathcal{I}(\boldsymbol{x};\boldsymbol{z})$.*