Appendix: Proving Thm 1

In this appendix, we will prove Thm 1. We will start with a few lemmas:

Lemma 1.1. *Given assumptions 1 and 2 in Thm. 1, S_1 is a deterministic one-to-one mapping of U_1; so is S_2 to U_2.*

The proof is obvious and omitted.

Lemma 1.2. *Given all the assumptions in Thm. 1. Then exist a asymptotic global minimizer of Eq. (5) that statisfies:*
1.

$$\lim_{T\to\infty} \frac{1}{T} I(C_1; U_1) = 0 \tag{13}$$

where $I(\cdot;\cdot)$ denotes mutual information.

2.

$$\operatorname*{plim}_{T\to\infty} \hat{X}_{1\to 1} = X_1 \tag{14}$$

3.

$$\lim_{T\to\infty} \mathbb{E}[(\hat{X}_{1\to 1} - X_1)^2] + \lambda \mathbb{E}[(E_c(\hat{X}_{1\to 1}) - C_1)^2] = 0 \tag{15}$$

Proof. Define the following set

$$\mathcal{X} = \{x_1 : \log p_{X_1}(x_1|U_1) \le n - 1 = n^* - 1 + T^{2/3}\} \tag{16}$$

$\mathcal{X}$ characterizes the set of instances where the optimal code length is guaranteed to be smaller than n.

Denote $C_1 = E_c^*(X_1; T)$ as the the following coding scheme. When $X_1 \in \mathcal{X}$, C_1 is the optimal lossless code for $p_{X_1}(\cdot|u_1)$ (whose code length is smaller than n by Shannon's Coding Theorem) padded with 0 to length n. When $X_1 \notin \mathcal{X}$, C_1 is any random number of dimension n.

Denote an auxiliary random variable

$$A_1 = \mathbb{1}[X_1 \in \mathcal{X}] \tag{17}$$

where $\mathbb{1}[\cdot]$ denotes the indicator function.

When $A_1 = 1$, there is a one-to-one mapping from C_1 to X_1, so we have

$$H(C_1|U_1, A_1 = 1) = H(X_1|U_1, A_1 = 1) \tag{18}$$

On the other hand, define h_m as the capacity of each dimension of C_1, *i.e.*

$$h_{\mathrm{m}} = \max_{p_{C_{1i}}(\cdot)} H(C_{1i}) \tag{19}$$

Then, the information C_1 contains is limited by the number of dimensions it has, *i.e.*

$$\begin{aligned} H(C_1) &\le \sum_i H(C_{1i}) \le n h_{\mathrm{m}} \\ &\le n^* h_m + T^{2/3} h_m \\ &\le H(X_1|U_1) + 1 + T^{2/3} h_m \end{aligned} \tag{20}$$

where the second line is from assumption 3 of Thm. 1. The third line is from the Shannon's coding theorem.

Notice that A_1 is a function of X_1, and thus we have

$$\begin{aligned} H(X_1|U_1) &= H(X_1, A_1|U_1) \\ &= H(X_1|U_1, A_1) + H(A_1|U_1) \\ &\le H(X_1|U_1, A_1) + H(A_1) \\ &= H(X_1|U_1, A_1 = 1) p_{A_1}(1) \\ &\quad + H(X_1|U_1, A_1 = 0) p_{A_1}(0) + H(A_1) \\ &= H(C_1|U_1, A_1 = 1) p_{A_1}(1) \\ &\quad + H(X_1|U_1, A_1 = 0) p_{A_1}(0) + H(A_1) \\ &\le H(C_1|U_1, A_1 = 1) p_{A_1}(1) \\ &\quad + H(C_1|U_1, A_1 = 0) p_{A_1}(0) \\ &\quad + H(X_1|U_1, A_1 = 0) p_{A_1}(0) + H(A_1) \\ &= H(C_1|U_1, A_1) \\ &\quad + H(X_1|U_1, A_1 = 0) p_{A_1}(0) + H(A_1) \\ &\le H(C_1|U_1) \\ &\quad + H(X_1|U_1, A_1 = 0) p_{A_1}(0) + H(A_1) \end{aligned} \tag{21}$$

where the last but three line is given by Eq. (18).

Eqs. (20) and (21) imply that

$$\begin{aligned} I(C_1; U_1) &= H(C_1) - H(C_1|U_1) \\ &\le 1 + T^{2/3} h_m \\ &\quad + H(X_1|U_1, A_1 = 0) p_{A_1}(0) + H(A_1) \end{aligned} \tag{22}$$

For any $t \le T$, $X_1(t)$ is a discrete random variable with finite support cardinality, denoted as K. Then we have

$$\begin{aligned} H(X_1|U_1, A_1 = 0) &\le \sum_{t=1}^{T} H(X_1(t)|U_1, A_1 = 0) \\ &\le T \log K \end{aligned} \tag{23}$$

On the other hand, notice that $\{X_1(t)\}$ is a stationary Markov process of order τ. We have

$$\begin{aligned} \log p_{X_1}(\cdot|U_1) = &\sum_{t=1}^{\tau} \log p_{X_1(t)}(\cdot|U_1, X_1(1 : t-1)) \\ &+ \sum_{t=\tau+1}^{T} \log p_{X_1(t)}(\cdot|U_1, X_1(t-\tau : t-1)) \end{aligned} \tag{24}$$

From the central limit theorem for ergodic Markov process

$$\lim_{T\to\infty} p_{A_1}(1) = 1, \quad \lim_{T\to\infty} p_{A_1}(0) = 0 \tag{25}$$

Combining Eqs. (22), (23) and (25), we have

$$\frac{1}{T}I(C_1;U_1) \leq \frac{1}{T}(1+T^{2/3}h_m + p_{A_1}(0)T\log K + H(A_1)) \\ \rightarrow 0, \text{ as } T\rightarrow\infty \tag{26}$$

Hence Eq. (13) is proved.

Next, for $X_1 \in \mathcal{X}$, notice that $[C_1, S_1]$ is a lossless code of $[X_1, U_1]$, because

$$\begin{aligned} H(X_1, U_1) &= H(U_1) + H(X_1|U_1) \\ &= H(U_1) + H(C_1|U_1) \\ &= H(S_1) + H(C_1|S_1) \\ &= H(C_1, S_1) \end{aligned} \tag{27}$$

where the second line is from Eq (18); the third line is from Lem. 1.1. Eq. (27) implies that $[U_1, X_1]$ is fully recoverable from $[C_1, S_1]$. Therefore, there exists an optimum decoder $D^*(\cdot,\cdot)$ such that

$$\hat{X}_{1\rightarrow 1} = X_1 \tag{28}$$

Combining Eqs. (25) and (28), Eq. (14) is proved.

Apply $E_c(\cdot)$ to both sides, we get

$$\operatorname*{plim}_{T\rightarrow\infty} E_c(\hat{X}_{1\rightarrow 1}) = E_c(X_1) = C_1 \tag{29}$$

Hence, considering X_1 has finite second moment, convergence with probability implies mean squared convergence, *i.e.*

$$\lim_{T\rightarrow\infty} \mathbb{E}[\|\hat{X}_{1\rightarrow 1} - X_1)^2\|_2^2] + \lambda\mathbb{E}[\|E_c(\hat{X}_{1\rightarrow 1}) - C_1\|_1] = 0 \tag{30}$$

which means that $[E_c^*(\cdot), D^*(\cdot,\cdot)]$ is the asymptotic global optimizer of Eq. (5). □

Now we are ready to prove Thm 1.

Proof. (Thm. 1) Denote X_2' as speech drawn from the ground truth distribution of the converted speech, *i.e.* $p_X(\cdot|U = U_2, Z = Z_1)$. Then our goal is to show that $\hat{X}_{1\rightarrow 2}$ is assymptotically identically distributed to X_2'.

What we will do is bridge the two random variables by passing X_2' to AUTOVC for self-reconstruction. Namely,

$$C_2' = E_c^*(X_2'), \text{ and } \hat{X}_{2\rightarrow 2}' = D^*(C_2', S_2) \tag{31}$$

where $E^*(\cdot)$ and $D(\cdot)$ are the optimal encoder and decoder derived in Lem. 1.2.

From Lem. 1.2, we know that $\hat{X}_{2\rightarrow 2}' \rightarrow X_2'$ with probability. So all is left to do is to show that $\hat{X}_{2\rightarrow 2}'$ is assymptotically identically distributed to $\hat{X}_{1\rightarrow 2}$.

First, notice that

$$\begin{aligned} p_{C_1}(\cdot|z_1, u_2) &= p_{C_1}(\cdot|z_1) \\ &= p_{E_c(X_1)}(\cdot|z_1) \\ &= p_{E_c(X)}(\cdot|Z = z_1) \end{aligned} \tag{32}$$

where the first line is due to the fact that C_1 and Z_1 are both independent of U_2 (Recall U_2 is not involved in the generation process of C_1); the last line is from the fact that (U_1, Z_1, X_1) is identically distributed to (U, Z, X).

Therefore, we can show that

$$\begin{aligned} &\lim_{T\rightarrow\infty}\frac{1}{T}KL(p_{C_2'}(\cdot|z_1,u_2)||p_{E_c(X)}(\cdot|Z=z_1)) \\ =&\lim_{T\rightarrow\infty}\frac{1}{T}KL(p_{C_2'}(\cdot|z_1,u_2)||p_{C_1}(\cdot|Z=z_1)) \\ =&0 \end{aligned} \tag{33}$$

where the last line is given by Eq. (13) of Lem. 1.2.

On the other hand,

$$\begin{aligned} p_{\hat{X}_{1\rightarrow 2}}(\cdot|z_1,u_2) &= p_{D^*(C_1,S_2)}(\cdot|z_1,u_2) \\ p_{\hat{X}_{2\rightarrow 2}'}(\cdot|z_1,u_2) &= p_{D^*(C_2',S_2)}(\cdot|z_1,u_2) \end{aligned} \tag{34}$$

Combining Eqs. (33) and (34), we have

$$\lim_{T\rightarrow\infty}\frac{1}{T}KL(p_{\hat{X}_{1\rightarrow 2}}(\cdot|z_1,u_2)||p_{\hat{X}_{2\rightarrow 2}'}(\cdot|z_1,u_2)) = 0 \tag{35}$$

□

Here is a final note on Thm 1. The content loss can help to constrain information capacity of the bottleneck by soft-constraining the range of each dimension of the content code, otherwise the information capacity of each bottleneck dimension can be unbounded and Thm 1 does not apply.