

# Supplementary File to "Seeking the Shape of Sound: An Adaptive Framework for Learning Voice-Face Association"

Anonymous CVPR 2021 submission

Paper ID 2586

## 1. Content Overview

In this supplementary file, we provide more details of our paper and qualitative results. This file is organized as follows:

- in Sec. 2, we provide proofs of two importance deductions in our paper.
- in Sec. 3, we provide more details on how we generate the queries used for evaluation.
- in Sec. 4, we provide some qualitative results to demonstrate the effect of the proposed method.

The code is under the folder *code\_release*, and will be released upon this paper is published.

## 2. Details of Derivation

In this section, we will provide the proof that adopting a common classifier could lead to an implicit modality alignment (**Proposition 1**), and that positive and negative imbalance will not occur in the explicit modality alignment (**Equation (5)**).

### 2.1. Proof of Proposition 1

**Restate of Proposition 1.** *Supposing that, for any  $k \in \{1, 2, \dots, M\}$ , the weight decay strategy ensures  $\|\omega_k\| \leq C$ , we have a lower bound of  $\mathcal{L}_{implicit}$  written as follows:*

$$\mathcal{L}_{implicit} \geq 2 \log M - \frac{C}{MN} \sum_{j=1}^M D_j$$

where

$$D_j = \left\| (M-1) \sum_{y_i=j} (\mathbf{x}_i + \mathbf{v}_i) - \sum_{y_i \neq j} (\mathbf{x}_i + \mathbf{v}_i) \right\|.$$

*Proof.*

$$\begin{aligned} \mathcal{L}_{face} &= \frac{1}{N} \sum_{i=1}^N \log \left[ \sum_{j=1}^M \exp((\omega_j^T - \omega_{y_i}^T) \mathbf{x}_i) \right] \\ &\stackrel{(1)}{\geq} \frac{1}{NM} \sum_{i=1}^N \sum_{j=1}^M \log [M \exp((\omega_j^T - \omega_{y_i}^T) \mathbf{x}_i)] \\ &= \log M + \frac{1}{MN} \sum_{i=1}^N \sum_{j=1}^M (\omega_j^T - \omega_{y_i}^T) \mathbf{x}_i \\ &= \log M + \frac{1}{MN} \left( \sum_{j=1}^M \omega_j^T \sum_{i=1}^N \mathbf{x}_i - M \sum_{j=1}^M \omega_{y_i}^T \sum_{y_i=j} \mathbf{x}_i \right) \\ &= \log M - \frac{1}{MN} \sum_{j=1}^M \omega_j^T \left( M \sum_{y_i=j} \mathbf{x}_i - \sum_{i=1}^N \mathbf{x}_i \right) \\ &= \log M - \frac{1}{MN} \sum_{j=1}^M \omega_j^T \left( (M-1) \sum_{y_i=j} \mathbf{x}_i - \sum_{y_i \neq j} \mathbf{x}_i \right) \end{aligned}$$

For the same reason:

$$\mathcal{L}_{voice} \geq \log M - \frac{1}{MN} \sum_{j=1}^M \omega_j^T \left( (M-1) \sum_{y_i=j} \mathbf{v}_i - \sum_{y_i \neq j} \mathbf{v}_i \right)$$

Therefore, we have:

$$\begin{aligned} \mathcal{L}_{implicit} &= \mathcal{L}_{voice} + \mathcal{L}_{face} \\ &\geq 2 \log M - \frac{1}{MN} \sum_{j=1}^M \omega_j^T \left[ (M-1) \sum_{y_i=j} (\mathbf{x}_i + \mathbf{v}_i) - \sum_{y_i \neq j} (\mathbf{x}_i + \mathbf{v}_i) \right] \\ &\stackrel{(2)}{\geq} 2 \log M - \frac{C}{MN} \sum_{j=1}^M D_j \end{aligned}$$

where

$$D_j = \left\| (M-1) \sum_{y_i=j} (\mathbf{x}_i + \mathbf{v}_i) - \sum_{y_i \neq j} (\mathbf{x}_i + \mathbf{v}_i) \right\|$$

- (1). It's due to the fact that the Jensen inequity, and log function is concave.
- (2). It's due to the fact that  $\mathbf{a}^T \mathbf{b} \leq \|\mathbf{a}\| \|\mathbf{b}\|$  and  $\|\omega_j\| \leq C$ .  $\square$

Similarly, after applying the identity re-weighting strategy (See Eq. (8), Eq. (9)), there is the following corollary:

**Corollary 1.** If  $\mathcal{L}_{implicit}$  is weighted with  $\hat{s}_i \in (0, 1)$ , i.e.,

$$\begin{aligned} \mathcal{L}_{face} &= - \sum_{i=1}^N \hat{s}_{y_i}^t \log \frac{\exp(\omega_{y_i} \mathbf{x}_i)}{\sum_{j=1}^M \exp(\omega_j \mathbf{x}_i)} \\ \mathcal{L}_{voice} &= - \sum_{i=1}^N \hat{s}_{y_i}^t \log \frac{\exp(\omega_{y_i} \mathbf{v}_i)}{\sum_{j=1}^M \exp(\omega_j \mathbf{v}_i)} \end{aligned}$$

we have a lower bound of  $\mathcal{L}_{implicit}$  written as follows:

$$\mathcal{L}_{implicit} \geq 2 \log M \frac{\sum_{i=1}^N \hat{s}_{y_i}}{N} - \frac{C}{MN} \sum_{j=1}^M \hat{D}_j$$

where

$$\hat{D}_j = \left\| (M-1) \sum_{y_i=j} \hat{s}_{y_i} (\mathbf{x}_i + \mathbf{v}_i) - \sum_{y_i \neq j} \hat{s}_{y_i} (\mathbf{x}_i + \mathbf{v}_i) \right\|.$$

Corollary 1 shows that after applying the re-weighting strategy,  $\mathcal{L}_{implicit}$  still has similar properties to the original one, except that minimizing  $\mathcal{L}_{implicit}$  leads to a larger value of  $\sum_{j=1}^M \hat{D}_j$  instead of  $\sum_{j=1}^M D_j$ . The difference is that embeddings are weighted according to identities in  $\hat{D}_j$ , which enables the embedding learning to be aware of the diversity of learning difficulty.

## 2.2. Proof of Equation (5)

**Restate of Equation (5).**

$$\begin{aligned} \mathcal{L}_{explicit} &= \frac{1}{N} \sum_{i=1}^N \log(m + \frac{\sum_{y_j \neq y_i} \exp(\mathbf{v}_i \hat{\mathbf{x}}_j)}{\exp(\mathbf{v}_i \hat{\mathbf{x}}_i)}) \\ &\quad + \frac{1}{N} \sum_{i=1}^N \log(m + \frac{\sum_{y_j \neq y_i} \exp(\mathbf{x}_i \hat{\mathbf{v}}_j)}{\exp(\mathbf{x}_i \hat{\mathbf{v}}_i)}) \\ &\approx \frac{1}{N} \sum_{i=1}^N [\max_{y_j \neq y_i} \{\mathbf{v}_i \hat{\mathbf{x}}_j\} - \mathbf{v}_i \hat{\mathbf{x}}_i + m - 1]_+ \\ &\quad + \frac{1}{N} \sum_{i=1}^N [\max_{y_j \neq y_i} \{\mathbf{x}_i \hat{\mathbf{v}}_j\} - \mathbf{x}_i \hat{\mathbf{v}}_i + m - 1]_+ \end{aligned}$$

where  $[x]_+$  indicates  $\max(x, 0)$ .

*Proof.* We just need to prove that

$$\begin{aligned} &\frac{1}{N} \sum_{i=1}^N \log(m + \frac{\sum_{y_j \neq y_i} \exp(\mathbf{v}_i \hat{\mathbf{x}}_j)}{\exp(\mathbf{v}_i \hat{\mathbf{x}}_i)}) \\ &\approx \frac{1}{N} \sum_{i=1}^N [\max_{y_j \neq y_i} \{\mathbf{v}_i \hat{\mathbf{x}}_j\} - \mathbf{v}_i \hat{\mathbf{x}}_i + m - 1]_+ \end{aligned}$$

In fact,

$$\begin{aligned} &\frac{1}{N} \sum_{i=1}^N \log(m + \frac{\sum_{y_j \neq y_i} \exp(\mathbf{v}_i \hat{\mathbf{x}}_j)}{\exp(\mathbf{v}_i \hat{\mathbf{x}}_i)}) \\ &= \frac{1}{N} \sum_{i=1}^N \log[1 + \exp(\log(m-1 + \sum_{y_j \neq y_i} \exp(\mathbf{v}_i (\hat{\mathbf{x}}_j - \hat{\mathbf{x}}_i)))] \\ &\stackrel{(1)}{\approx} \frac{1}{N} \sum_{i=1}^N [\log(m-1 + \sum_{y_j \neq y_i} \exp(\mathbf{v}_i (\hat{\mathbf{x}}_j - \hat{\mathbf{x}}_i)))]_+ \\ &\stackrel{(2)}{\approx} \frac{1}{N} \sum_{i=1}^N [\max_{y_j \neq y_i} \{\mathbf{v}_i \hat{\mathbf{x}}_j - \mathbf{v}_i \hat{\mathbf{x}}_i\} + m - 1]_+ \\ &= \frac{1}{N} \sum_{i=1}^N [\max_{y_j \neq y_i} \{\mathbf{v}_i \hat{\mathbf{x}}_j\} - \mathbf{v}_i \hat{\mathbf{x}}_i + m - 1]_+ \end{aligned}$$

- (1). It's due to the fact that the softplus function  $\log(1 + \exp(x))$  approximates to hinge function  $[x]_+$ .
- (2). It's due to the fact that the logsumexp function  $\log(\gamma + \sum_i \exp(x_i))$  with a small constant  $\gamma$  approximates to the maximum of  $x_i + \gamma$ .  $\square$

## 3. Details of Query Generation in Datasets

In this section, we provide more details on how we generate the queries used to evaluate the model performance in the following four settings:

- (a) **1 : N matching.** In the 1 : N matching setting (where  $N = 2$  is a special case), for each instance in the test set, we randomly generate 3 galleries. A gallery consists of a randomly selected positive sample with the same identity of the probe and  $N - 1$  negative samples from other identities.
- (b) **Verification.** The verification testing queries are generated based on the 1 : 2 matching queries. For each probe, we randomly selected one sample from the gallery to query.
- (c) **Retrieval.** In this setting, we use all instances for testing. Each query contains a probe, and a gallery consists of all instances from the other modality. There are 21,850 instances of voice modality, and 20,076 instances of face modality.

the query lists are available in the folder *query\_lists*.

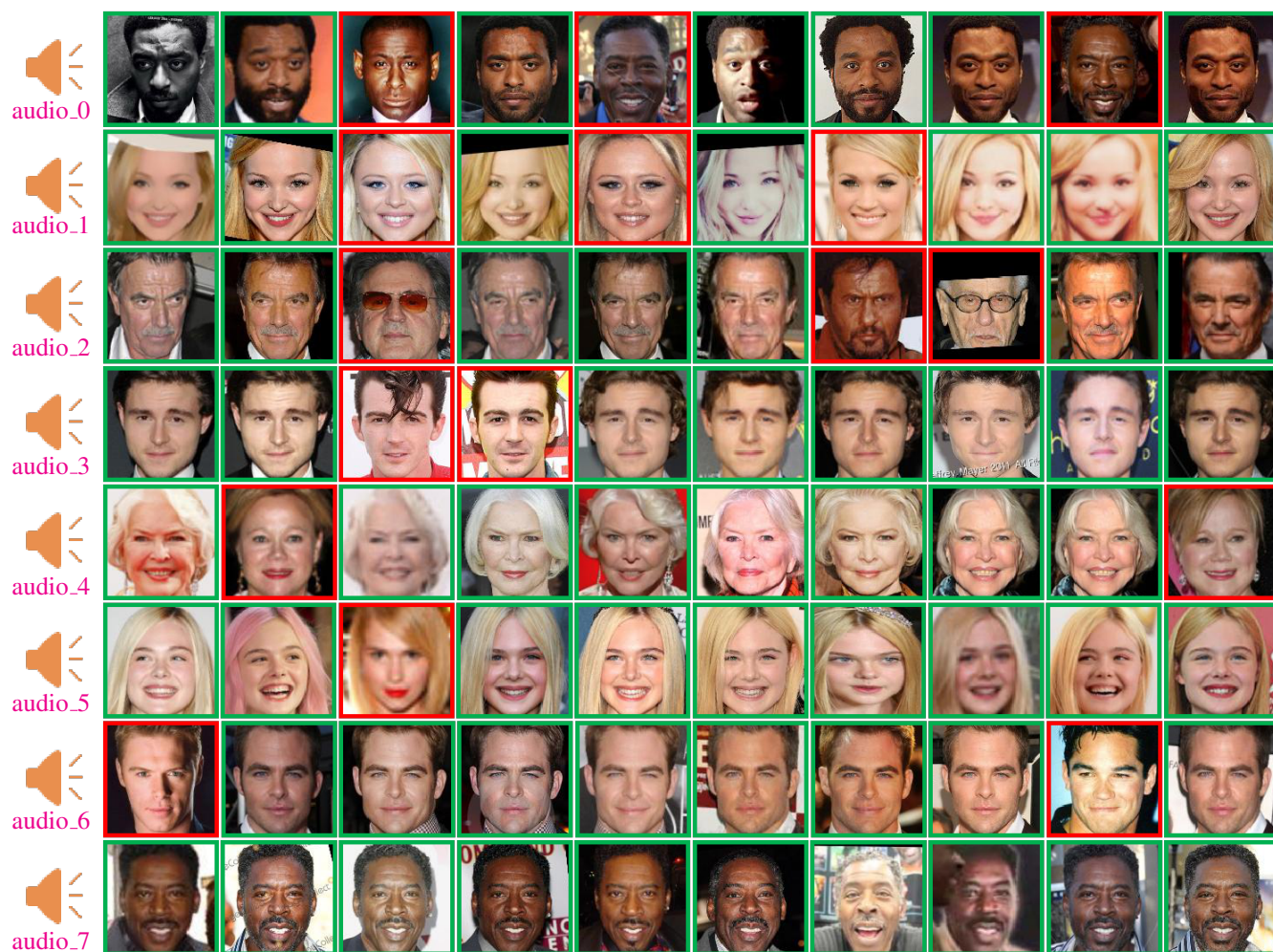


Figure 1. Qualitative results on voice to face retrieval. Top 10 of gallery are shown, where matched faces are marked in red, and mismatched faces are marked in green. Best viewed in color.

## 4. Qualitative Results

In the section, we provide some qualitative results under the most challenging setting, i.e., cross-modal retrieval,

in Fig. 1. For the convenience of demonstration, we only show results of retrieval from voice to face. Click on the text below the horn icon to play audio. The results are also integrated into a video file [retrieval\\_result.mp4](#).