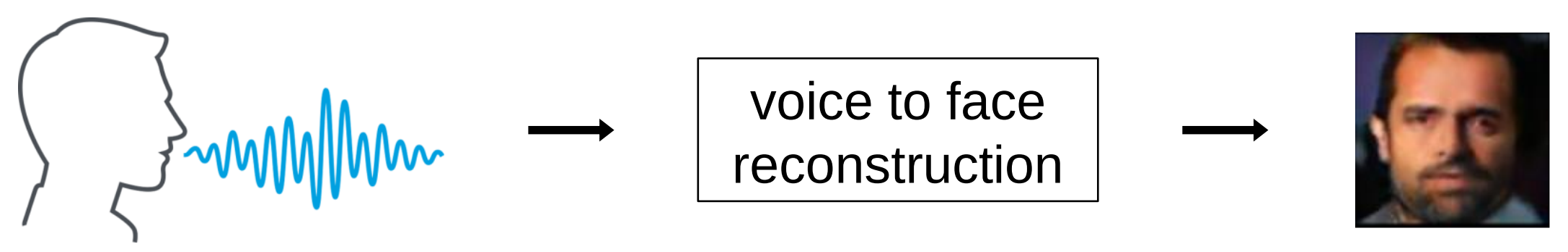
# Face Reconstruction from Voice using Generative Adversarial Networks

Yandong Wen, Rita Singh, Bhiksha Raj
Carnegie Mellon University

## Introduction

➢ **A New Task**

Given an audio clip spoken by an unseen person, we picture a face image that has as many associations as possible with the speaker , in terms of identity.


voice to face reconstruction

☒ make-up, expression, hair style, pose, etc.

☑ age, gender, ethnicity, etc.

➢ **Is it possible?**

- Skeletal and articulator structure of the face govern the shapes, sizes, and acoustic properties of the vocal tract that produces voice. []
- The same genetic physical and environmental influences that affect the development of the face also affect the voice. []
- Demographic factors influence both voice and face.

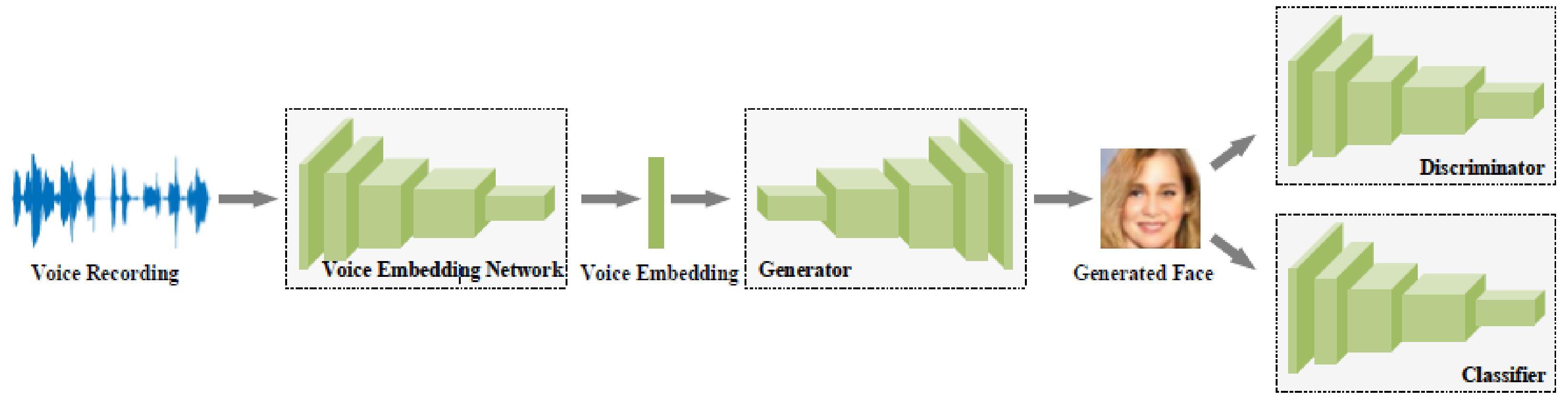
➢ **Challenges**

- It may not be able to entirely disambiguate all the face-related factors from the voice.
- It is unknown *a priori* exactly what features of the voice encode information about any given facial feature.
- It may not be sufficient for estimating a face image using the information containing in a single audio clip

➢ **Contributions**

- Introduce a new task of generating faces from voice in voice profiling.
- Propose a simple but effective framework based on generative adversarial networks.
- Propose to quantitatively evaluate the generated faces by using a cross-modal matching task.

## The proposed framework



- The voice-face correspondence is based on subject rather than sample.
- Paired voices and faces data are **NOT** required in each minibatch.

## Dataset

| | Train | Validation | Test | total |
|---|---|---|---|---|
| # of speech segments | 113,322 | 14,182 | 21,850 | 149,354 |
| # of face images | 106,584 | 12,533 | 20,455 | 139,572 |
| # of subjects | 924 | 112 | 189 | 1,225 |

**Table1**. Statistics of the Voxceleb1 dataset

## Qualitative Results

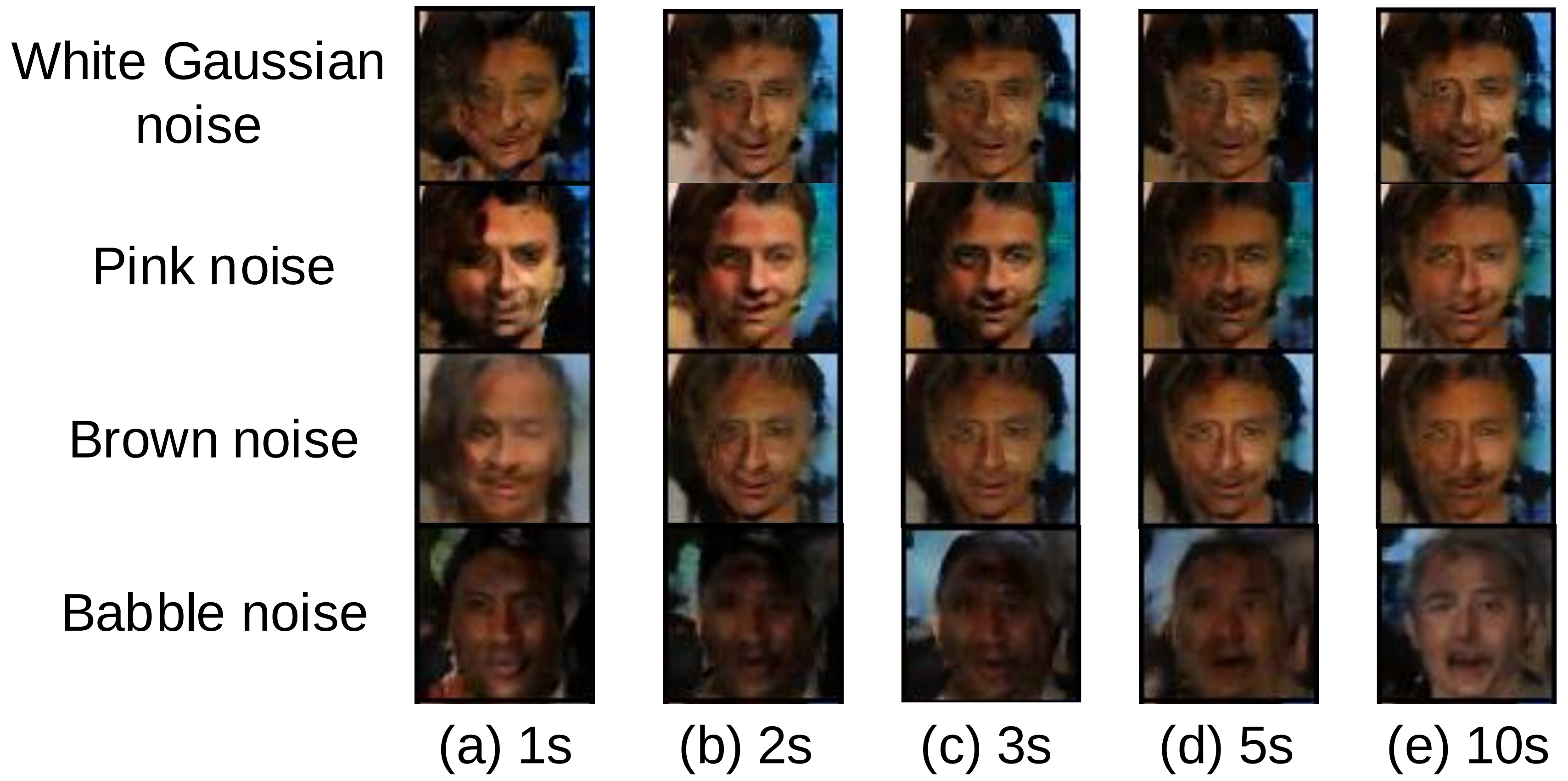


**Figure 1**. Each row shows the generated faces using one of the four noise audio segments with different durations.

## Qualitative Results (cont.)

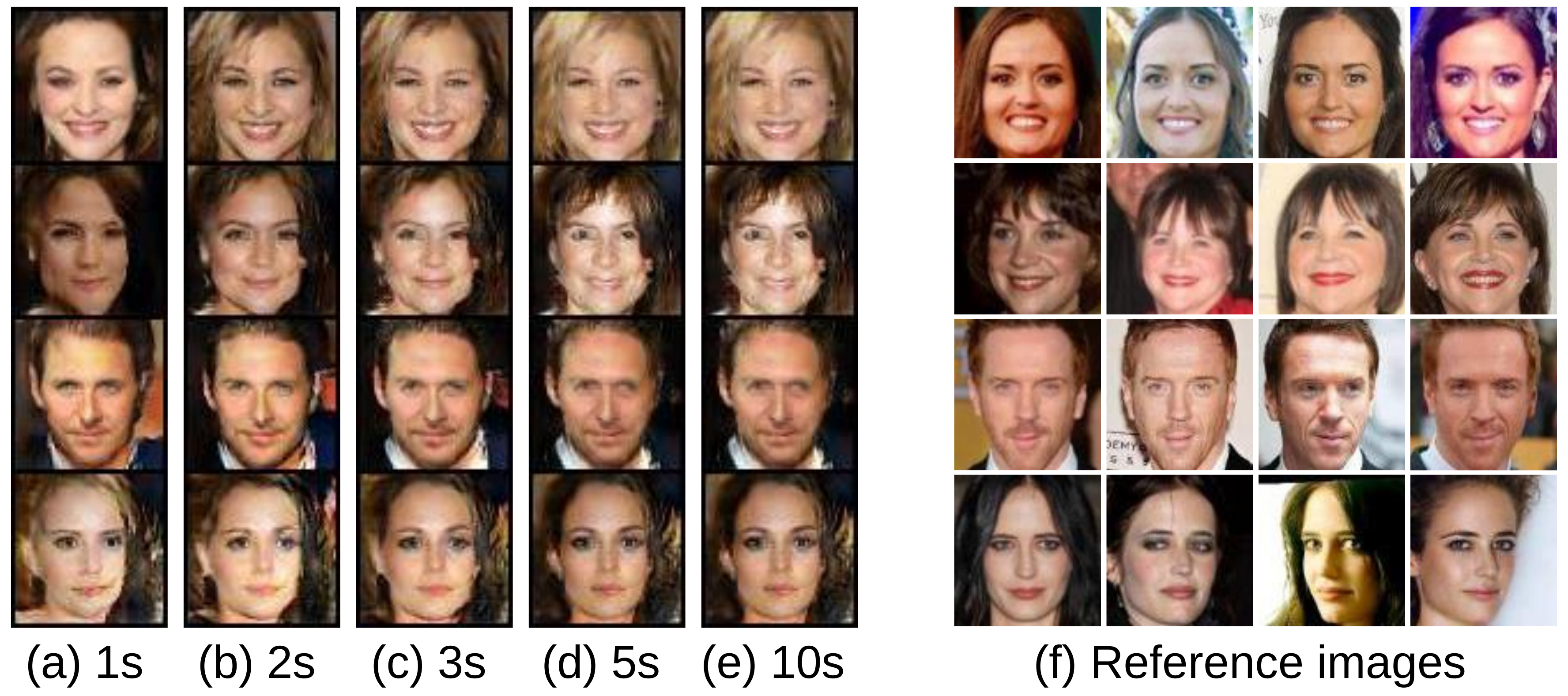


**Figure 2**. (a)-(e) The generated face images from regular speech recordings with different durations. (f) The reference face images

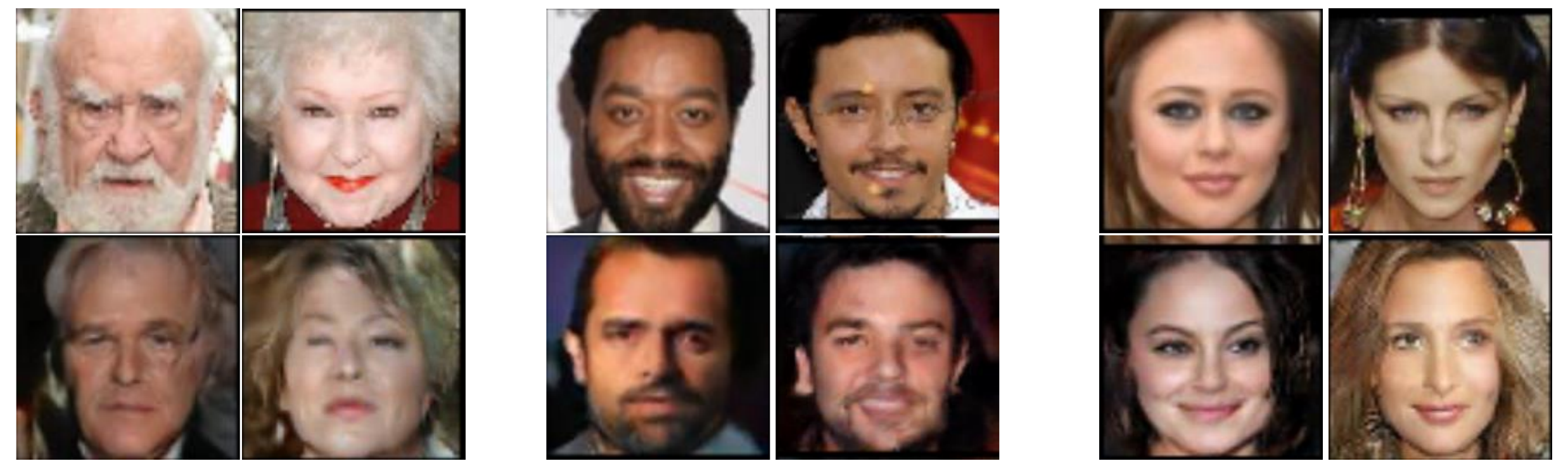


**Figure 3**. More examples with different genders and ages.

## Quantitative Results

| | unstratified group (ACC. %) (train / test) | stratified group by gender (ACC. %) (train / test) |
|---|---|---|
| SVHF | - / 81.00 | - / 65.20 |
| DIMNets-I | - / 83.45 | - / 70.91 |
| DIMNets-G | - / 72.90 | - / 50.32 |
| ours | 96.83/76.07 | 93.98 / 59.69 |

**Table 2**. The voice to face matching accuracies.

➢ Our results are given by replacing the probe voice embeddings by the embeddings of the generated face.