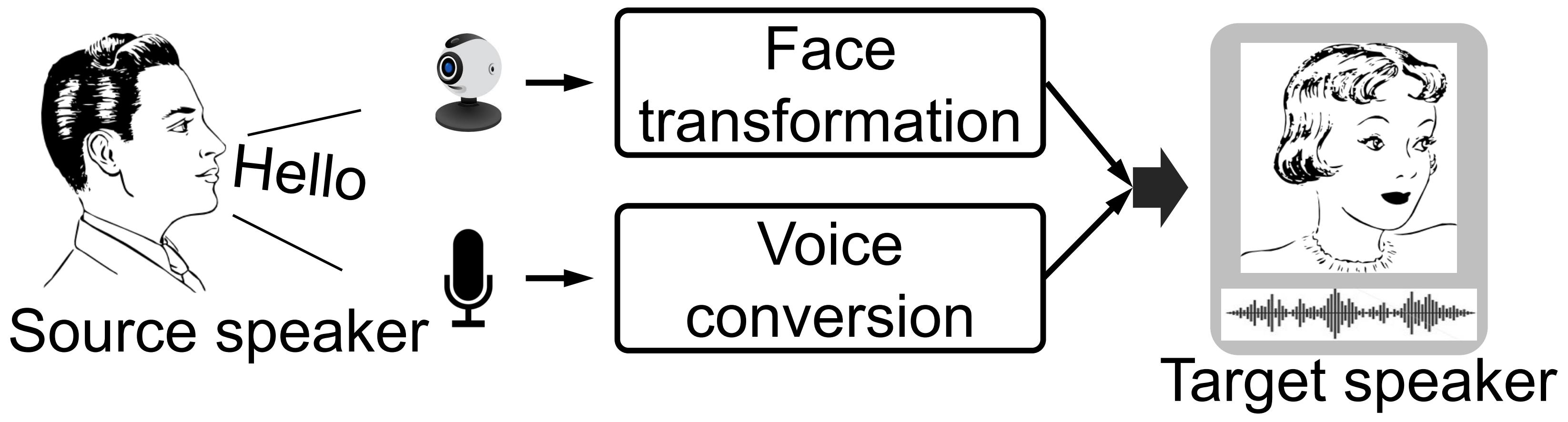
Audiovisual speaker conversion: jointly and simultaneously transforming facial expression and acoustic characteristics

Fuming Fang[1], Xin Wang[1], Junichi Yamagishi[1,2], Isao Echizen[1]
[1]National Institute of Informatics, [2]University of Edinburgh

1. Background

- Voice conversion: modifies acoustic individuality
- Face transformation: modifies facial individuality
- Applications: privacy protection, film production, games

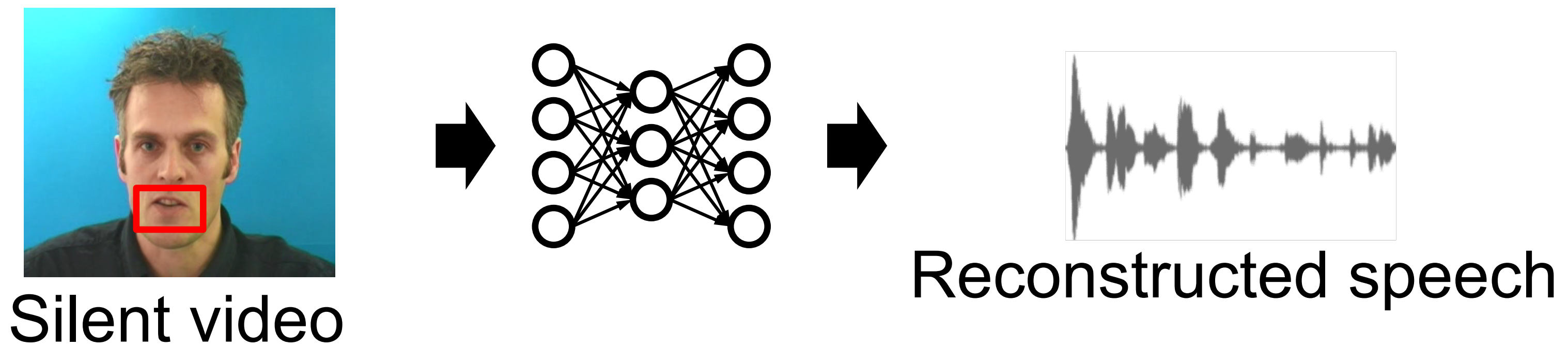


Separately transforming (existing methods)
⇒ Facial moving and speech are asynchronous

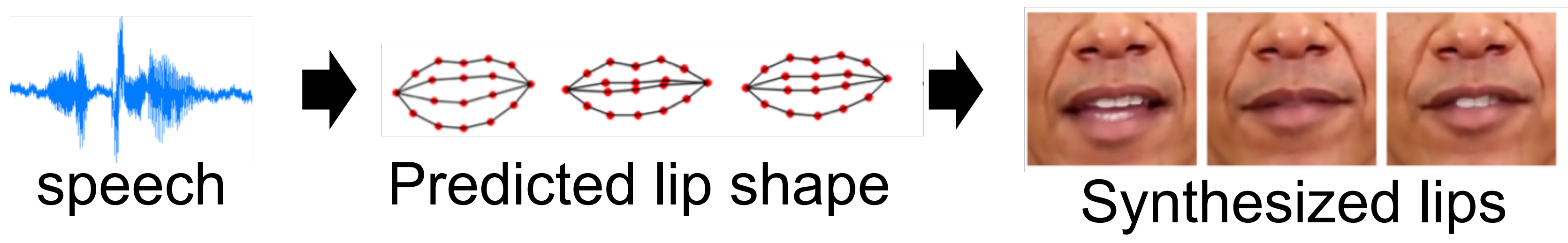
- This research: Jointly and simultaneously transforming
  ⇒ High correlation between facial moving and speech
  ⇒ Synchronous and more natural

2. Related works

- Audiovisual voice conversion [S. Tamura et al., '18]
  Audiovisual speech enhancement [T. Afouras et al. '18]
  - Using lip movements as extra information
- Lip moving-to-speech [Y. Kumar et al., '18]:



- Speech-to-lip moving [S. Suwajanakorn et al., '17]:

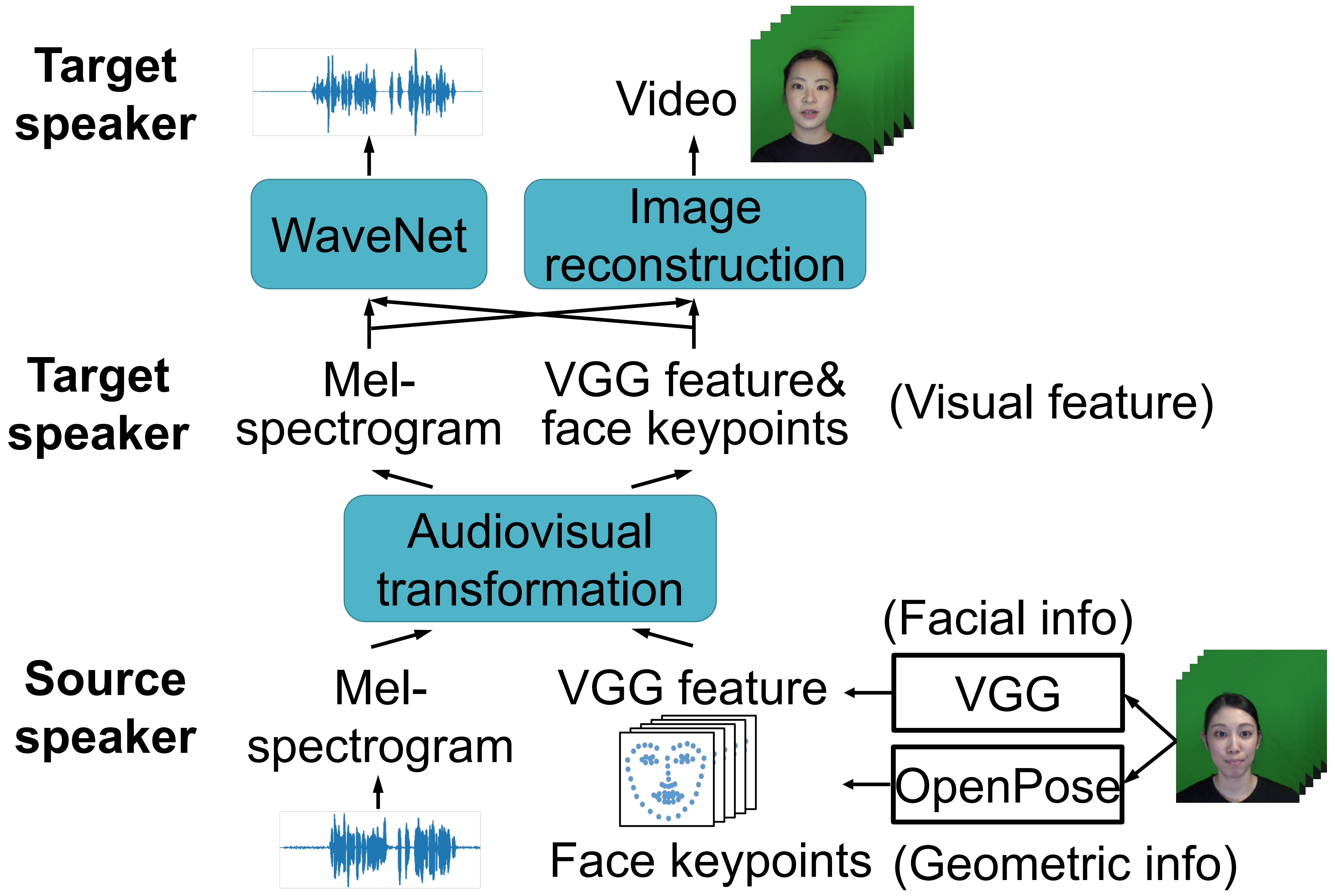


□ Our work differs from:

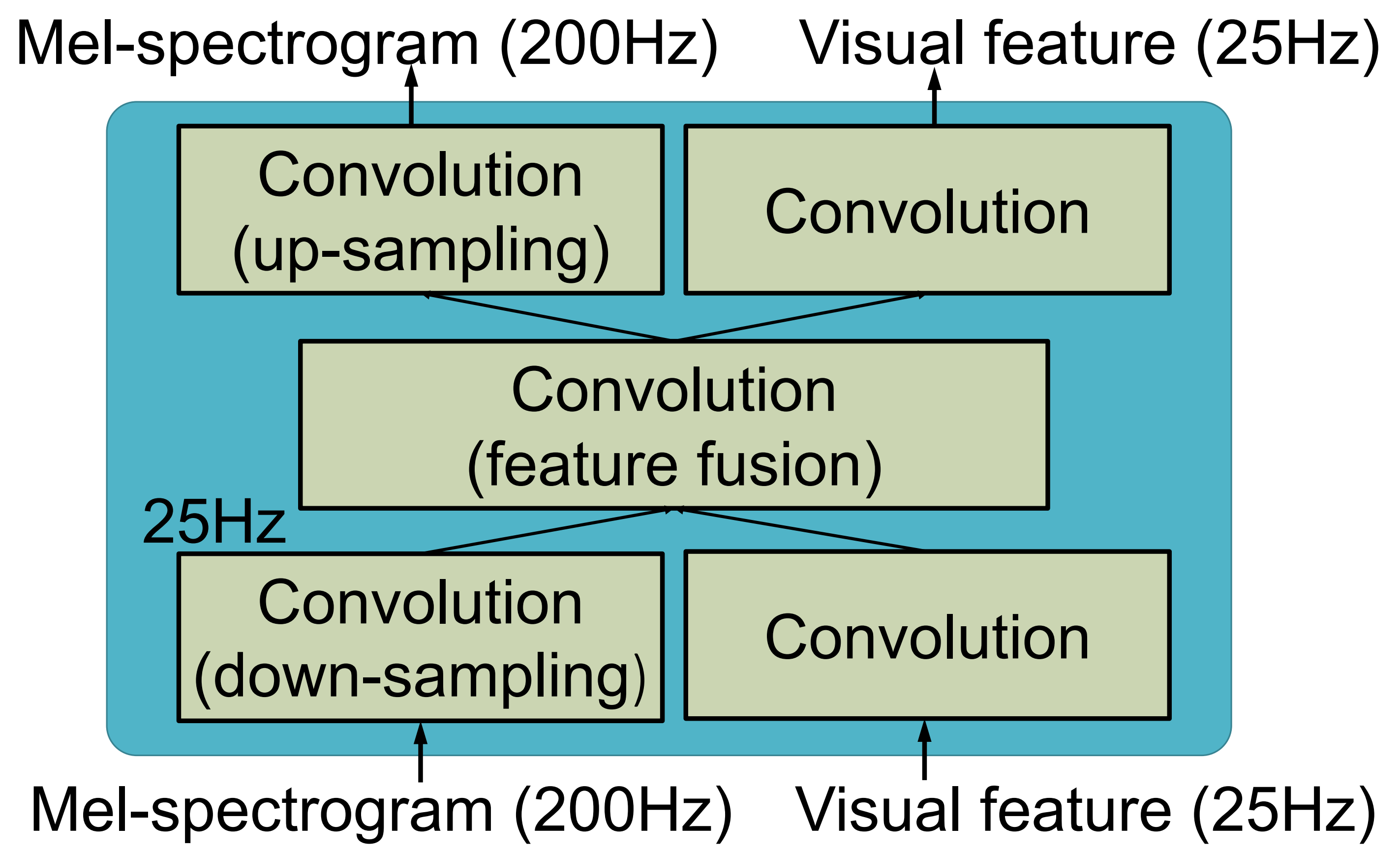
1. Using not only lip moving but also facial expression
2. Transforming both face and speech

3. Proposed method: audiovisual speaker conversion

- System consists of three networks
- Each network inputs both facial and acoustic features



- Audiovisual transformation network:
  - Performing convolution along with temporal axis



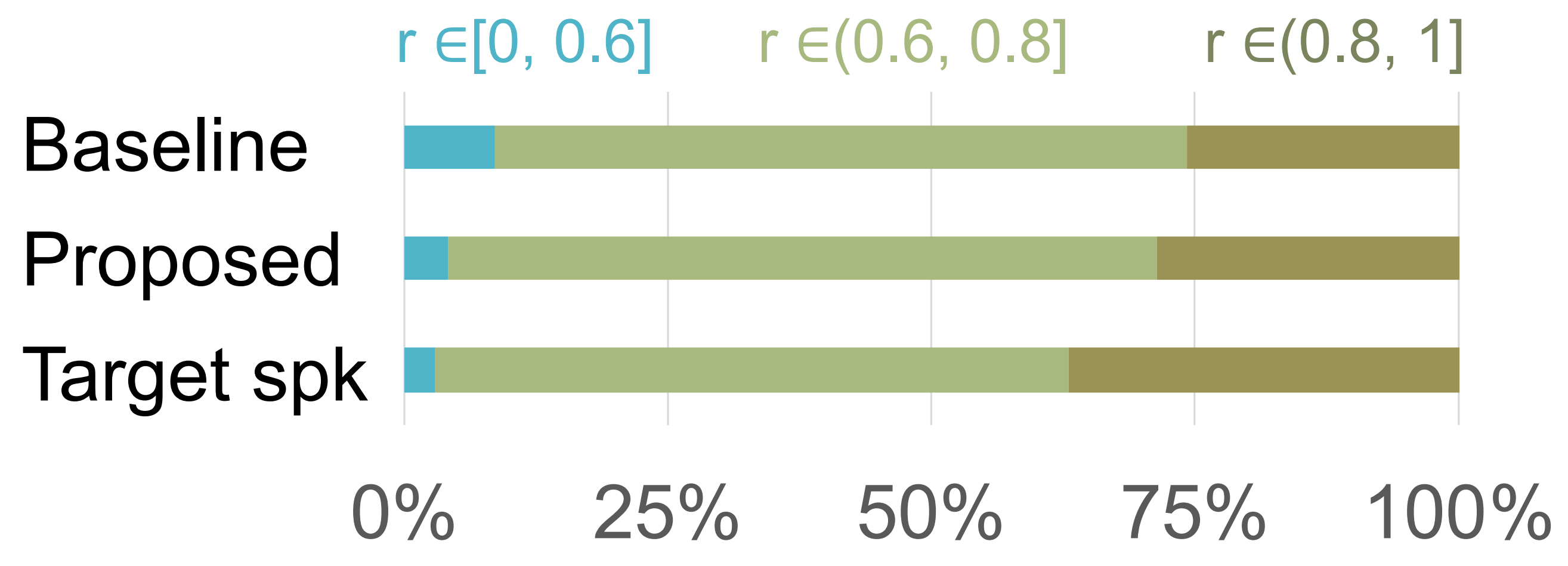
- Image reconstruction network:
  - LSGAN is used to train the network
  - Frame-to-frame mapping

4. Experiment

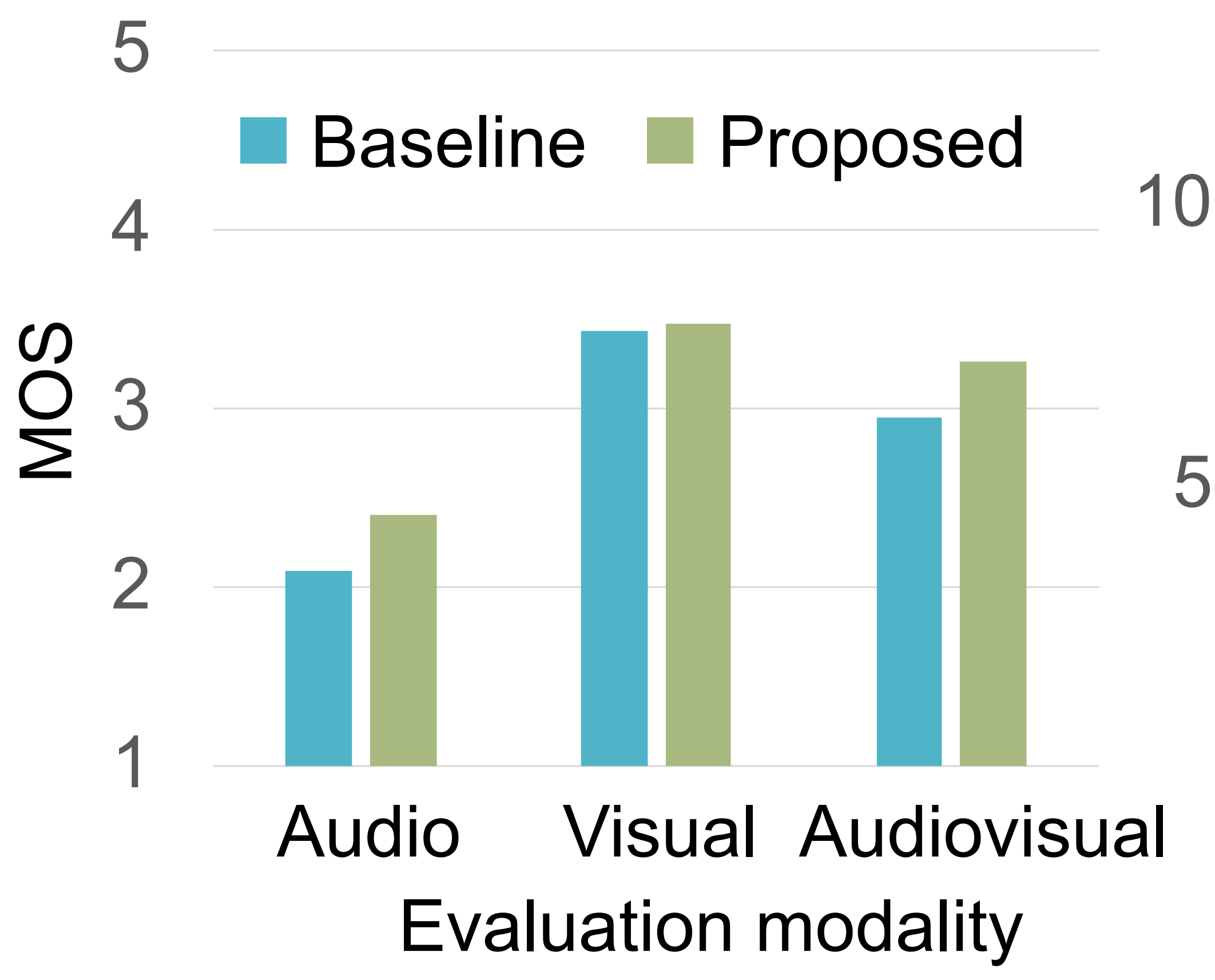
- Setup

Database	2 Japanese females
Mel-spectrum	80 dimensions
VGG feature	4096 dimensions
Face keypoints	140 dimensions
Subjective test	186 evaluators
Baseline	**Separated transformation**

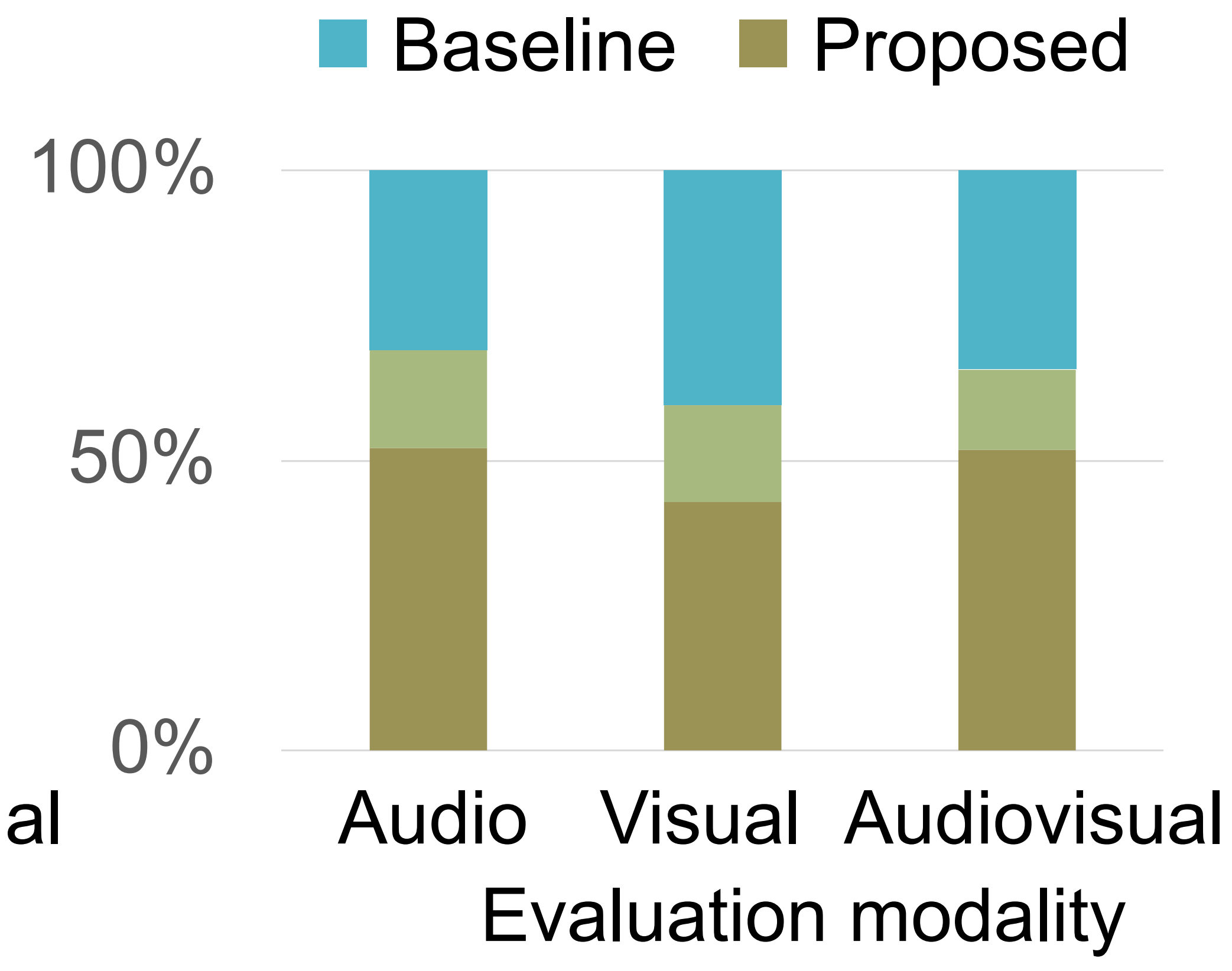
- Lip movements and speech correlation (r)



- Naturalness



- Speaker similarity (preference test)



- Higher naturalness because of higher correlation
- Visual feature dominated the conversion
- Difficult to balance visual and acoustic features

5. Conclusion & Future work

- Proposed an audiovisual speaker conversion method, by which facial and acoustic information can be highly correlated together
- Achieved higher naturalness compared to separated transformation
- Plan to reduce alignment error using CycleGAN-based non-parallel conversion