# Embedding vector
# in Multi-modality
# ( Audio signal processing )

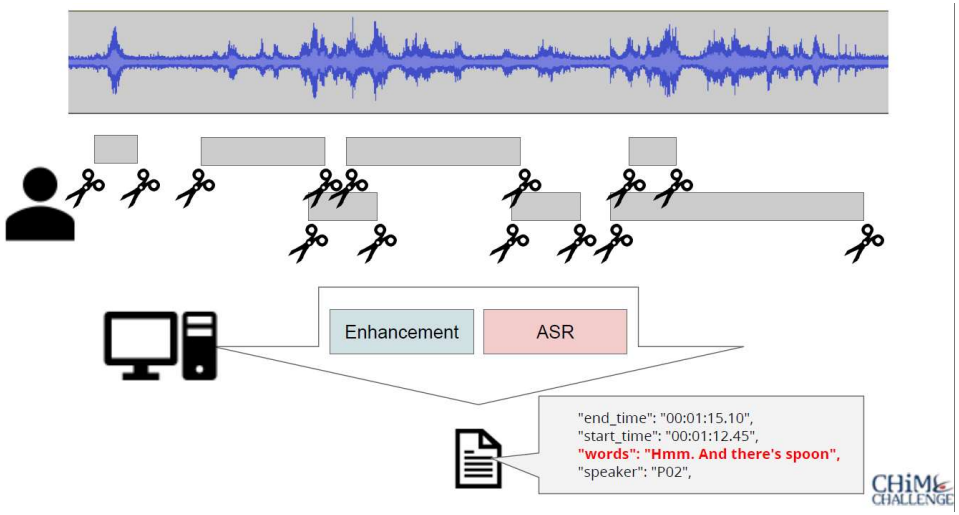
2020-11-24

문성규

# References


- D, Yu (Tencent), “Solving Cocktail Party Problem” CHiME 6, 2020
- D. Kitamura, “Audio Source Separation Based on Low-Rank Structure and Statistical Independence”
- D. Kitamura, “Super resolution-based stereo signal separation via supervised nonnegative matrix factorization”
- T. Kim (Neosapience), “Introduction to Speech Separation”

# Audio signal enhancement ?



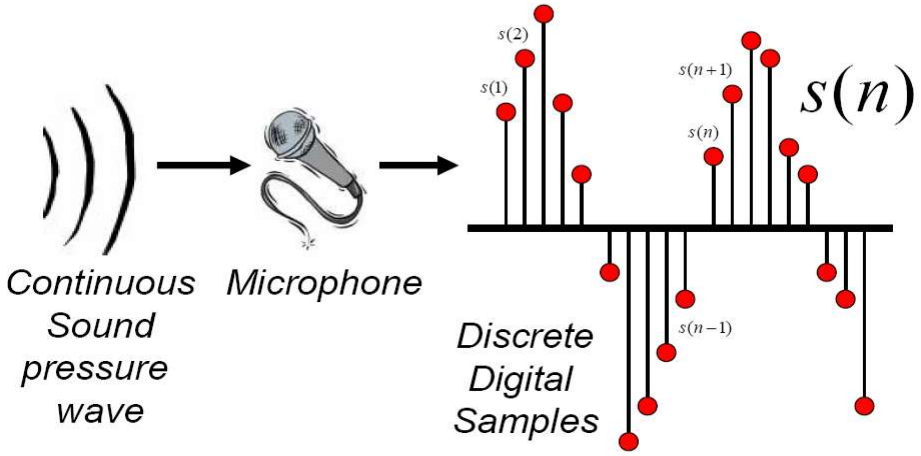

**CHiME-6 Challenge:**
**Tackling Multispeaker Speech Recognition for Unsegmented Recordings**

[1]*Shinji Watanabe,* [2]*Michael Mandel,* [3]*Jon Barker,* [4]*Emmanuel Vincent*
[1]*Ashish Arora,* [1]*Xuankai Chang,* [1]*Sanjeev Khudanpur,* [1]*Vimal Manohar,* [1]*Daniel Povey,* [1]*Desh Raj,*
[1]*David Snyder,* [1]*Aswin Shanmugam Subramanian,* [1]*Jan Trmal,* [1]*Bar Ben Yair,* [5]*Christoph Boeddeker,*
[2]*Zhaoheng Ni,* [6]*Yusuke Fujita,* [6]*Shota Horiguchi,* [7]*Naoyuki Kanda,* [7]*Takuya Yoshioka,* [8]*Neville Ryant*

[1]Johns Hopkins University, USA, [2]The City University of New York, USA, [3]University of Sheffield, UK, [4]Inria, France, [5]Paderborn University, Germany, [6]Hitachi, Ltd., Japan, [7]Microsoft, USA, [8]Linguistic Data Consortium, USA

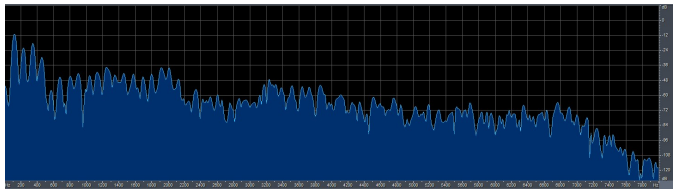
# Audio signal ?

- Continuous sound pressure → (sampled) discrete signal

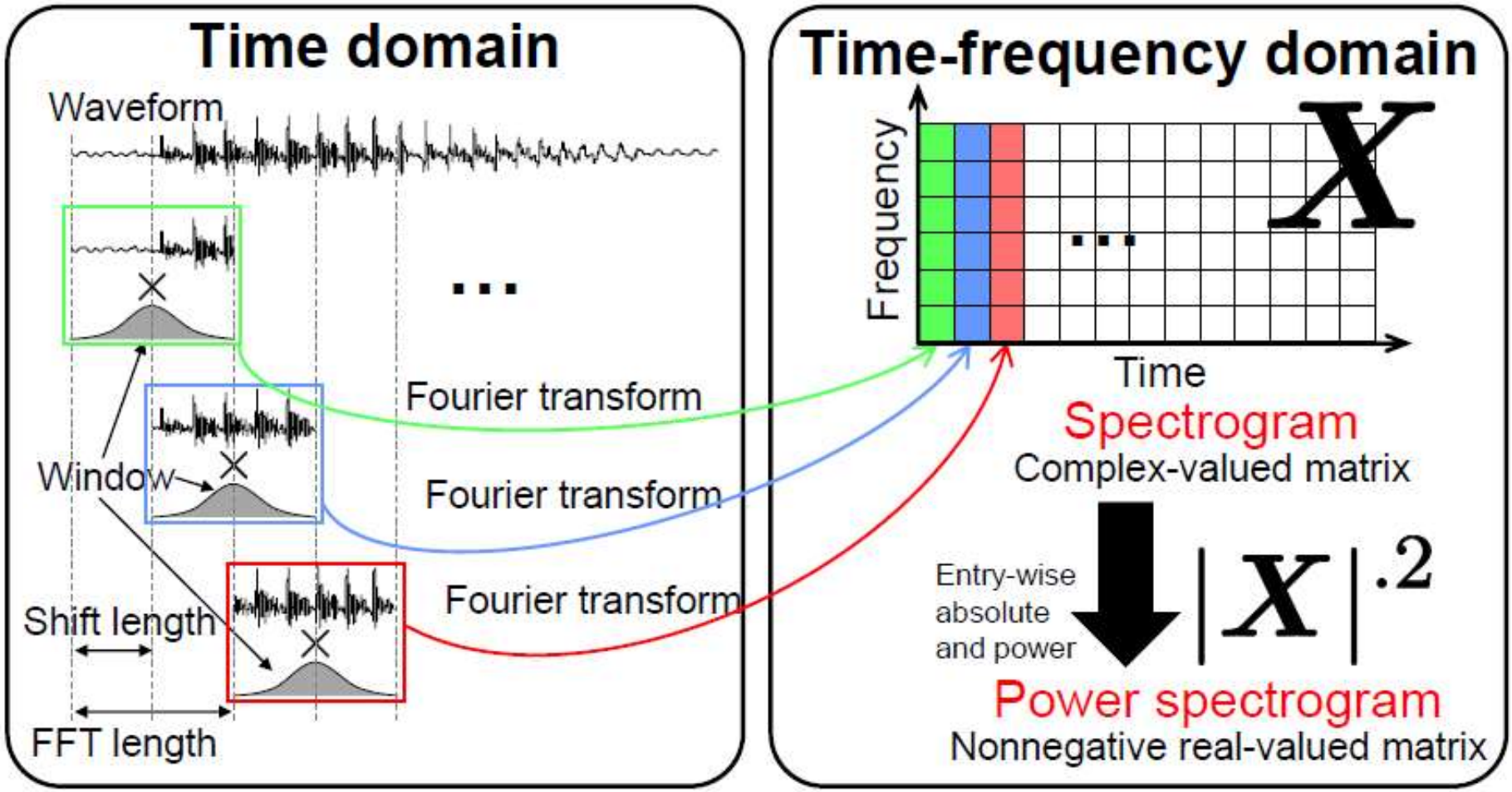



- Sampling:
  - Measuring amplitude of signal at time *t*
  - 44.1K Hz [samples/sec] : 가청 주파수 대역 (20KHz) 고려
  - 16,000 Hz : 일반적인 신호처리
  - 8,000 Hz : 전화 통화

# Audio signal ?



- Time-varying frequency structure
  - Short-time Fourier transform (STFT)

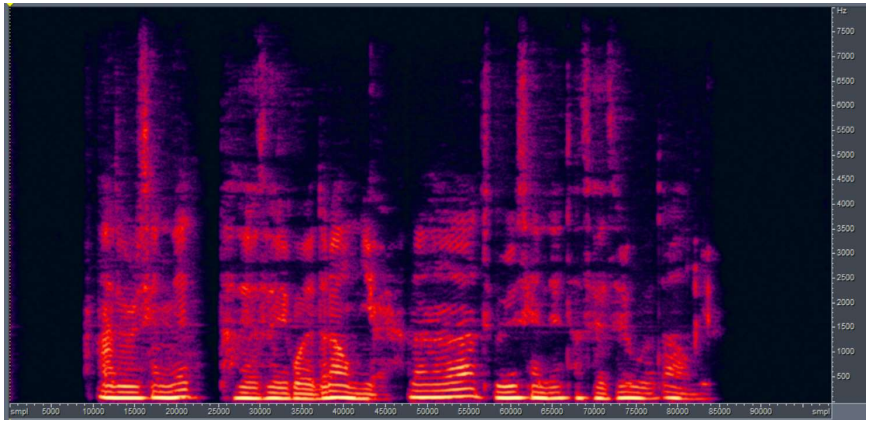



D. Kitamura, “Audio Source Separation Based on Low-Rank Structure and Statistical Independence”

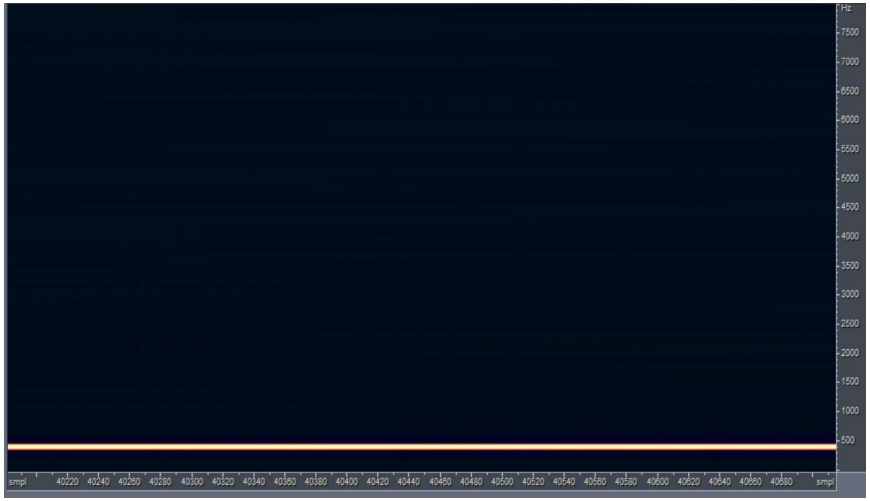
# Audio signal ?

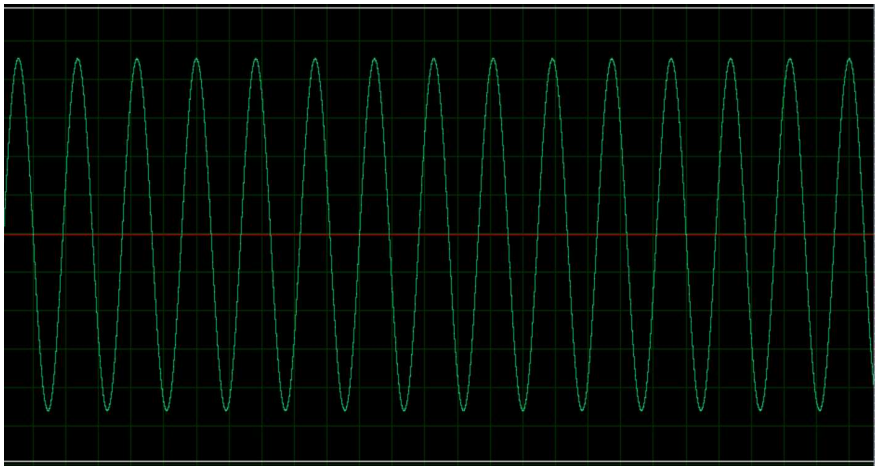
- 3Blue1Brown “Fourier series”
  https://www.youtube.com/watch?v=r6sGWTCMz2k

- Clean speech (Power spectrogram & Waveform)



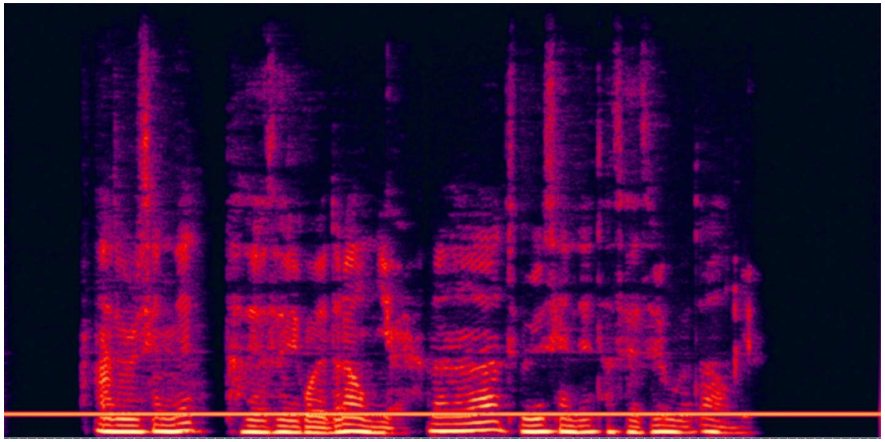
- 440Hz Sine wave

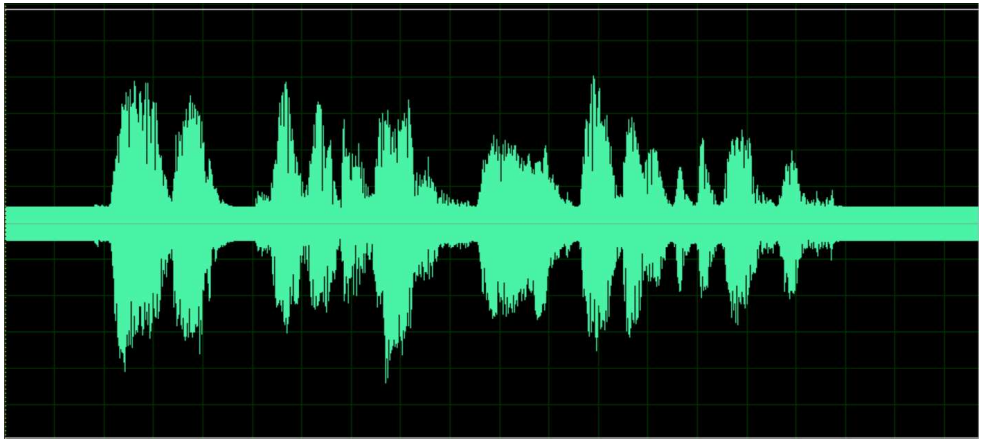




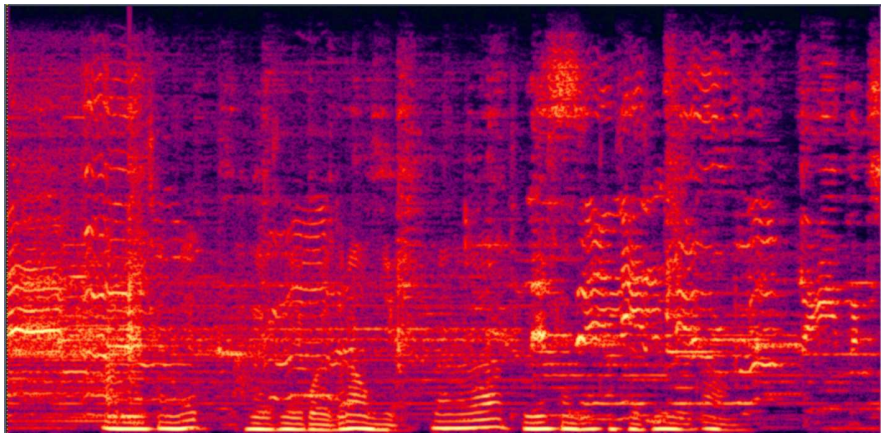
# Audio signal ?

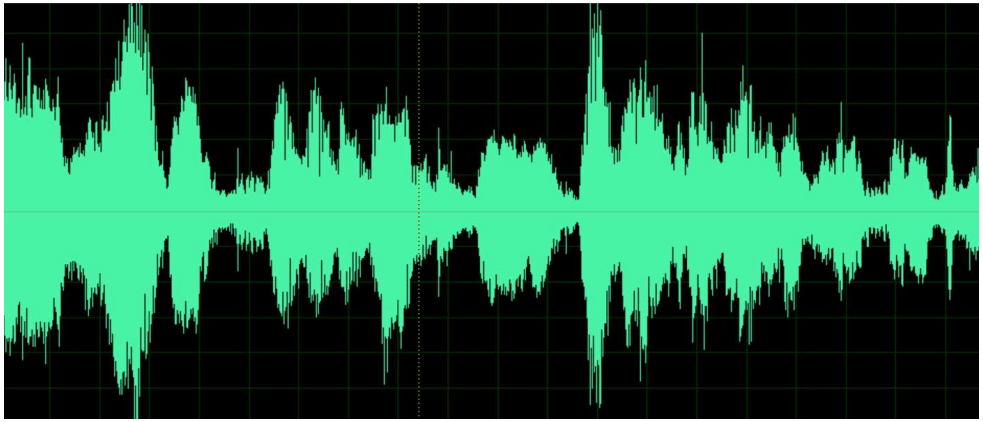
- Noisy speech (Speech + 440Hz Sine wave)





- Noisy speech (Speech + Interfering Speech or Music)





# Audio signal ?

▶항상 나오는 **질문...**

질문) speech recognition에선 frequncy domain feature인 MFCC로 바꾸는 preprocess를 하는데요, 왜 꼭 이 preprocess가 필요한건가요? 딥러닝의 모토가 end-to-end이고 이미지도 픽셀로부터 배우는데, 음성은 그냥 time domain 그대로 쓰면 안되는 (성능이 저하되는) 이유가 있을까요?

▶항상 나오는 **답변...**

**Learning the Speech Front-end With Raw Waveform CLDNNs**

*Tara N. Sainath, Ron J. Weiss, Andrew Senior, Kevin W. Wilson, Oriol Vinyals*

Google, Inc. New York, NY, U.S.A

{tsainath, ronw, andrewsenior, kwwilson, vinyals

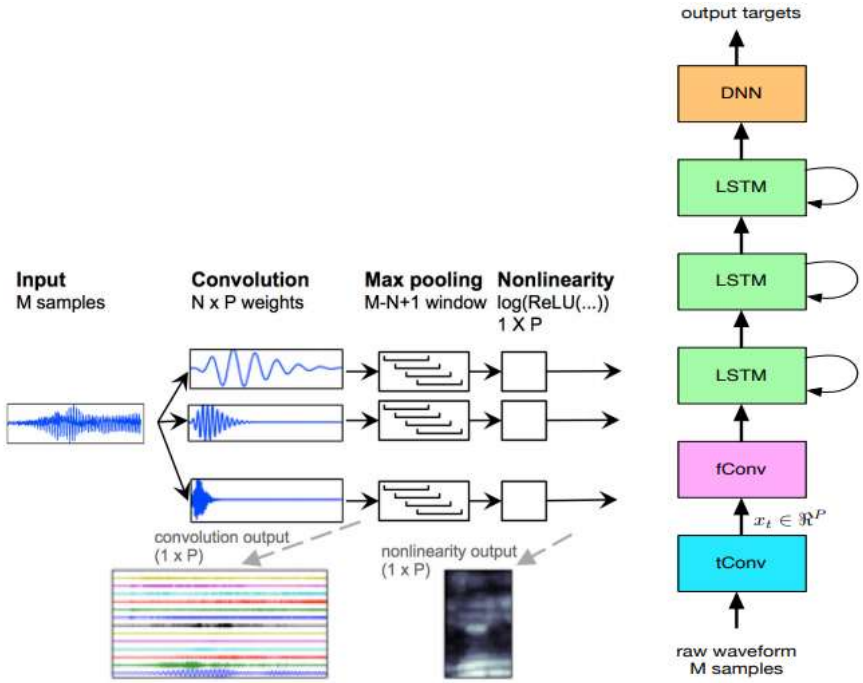
**Fully Convolutional Speech Recognition**

*Neil Zeghidour*[1,2,*], *Qiantong Xu*[1,*], *Vitaliy Liptchinsky*[1], *Nicolas Usunier*[1], *Gabriel Synnaeve*[1], *Ronan Collobert*[1]

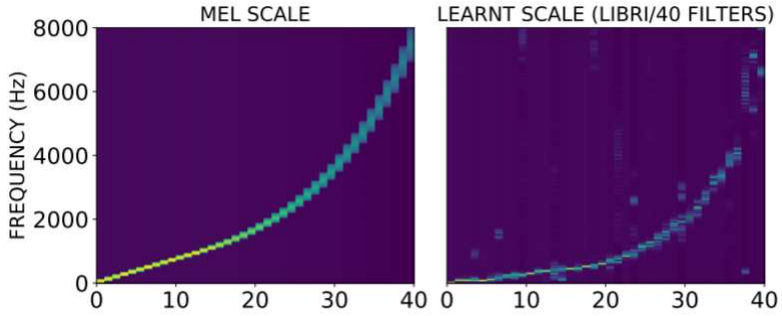
[1] Facebook A.I. Research, Paris, France; New York & Menlo Park, USA
[2] CoML, ENS/CNRS/EHESS/INRIA/PSL Research University, Paris, France

# Audio signal ?

▶항상 나오는 **답변...**




(a) Time-domain Convolution Layer (b) Time convolution and CLDNN Layers




| Feature | WER - CE | WER - Seq |
|---|---|---|
| raw | 16.2 | 14.2 |
| log-mel | 16.2 | 14.2 |
| raw+log-mel | **15.7** | **13.8** |

Table 6: WER Combining Raw and Log-Mel Features

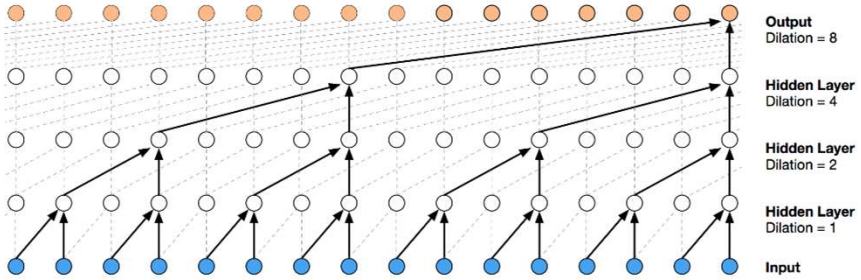
| Front-end | LM | | | | |
|---|---|---|---|---|---|
| Mel | 4-gram | 4.26 | 13.80 | 4.82 | 14.54 |
| Mel | ConvLM | 3.13 | 10.61 | 3.45 | 11.92 |
| Learnable (40) | ConvLM | 3.16 | 10.05 | 3.44 | 11.24 |
| Learnable (80) | ConvLM | 3.08 | 9.94 | 3.26 | 10.47 |

Table 2: *WER (%) on Librispeech.*

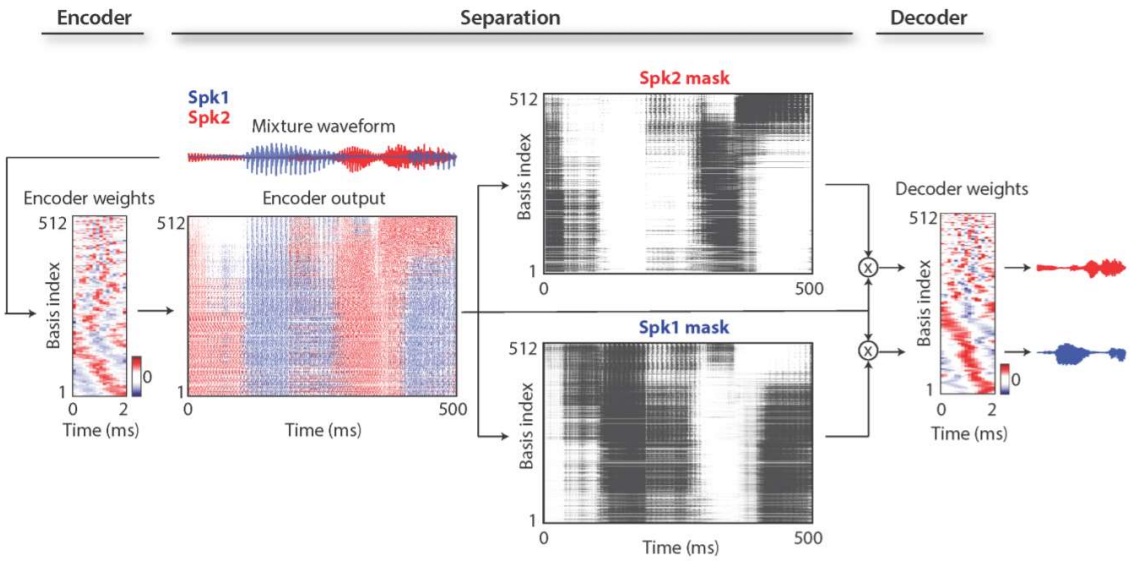
# Audio signal ?

▶항상 나오는 **반례...**

- A. Oord, 'Wavenet: A generative model for raw audio'

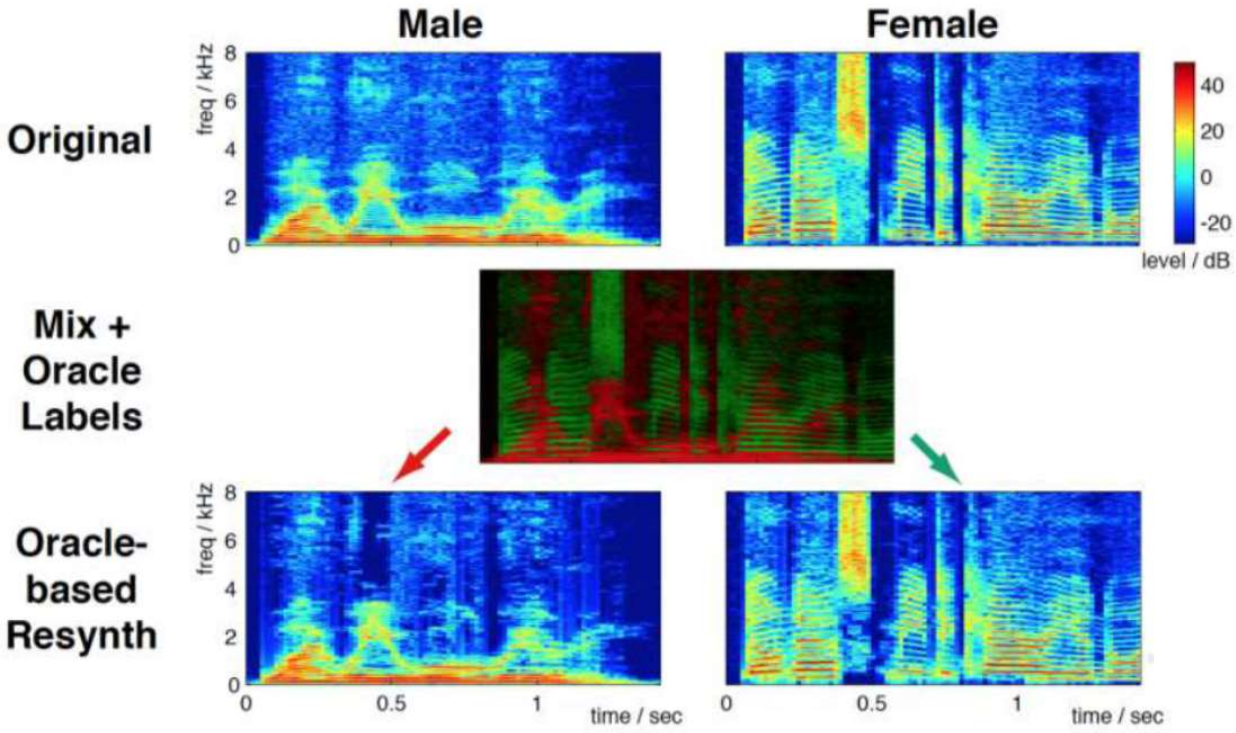



- Y. Luo, 'Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation'



# Audio signal enhancement

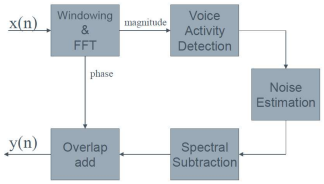
- T. Kim (Neosapience), “Introduction to Speech Separation”
  - https://www.youtube.com/watch?v=OgNSFKeHy8k




Spectral subtraction

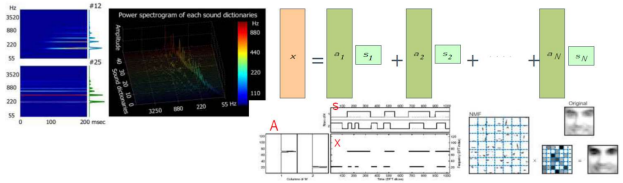
› Basic concept : |Y(f, t)| = |X(f, t)| – |N(f, t)|

› Procedure




Single channel approach in spectrogram

› Non-negative Matrix Factorization

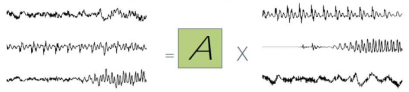


› Computational Auditory Scene Analysis
– Segmentation of auditory elements in spectrogram

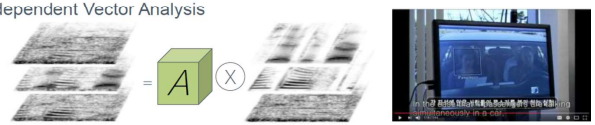
For more information about NMF approach in Korean
권기수, 김남수 (2016.5). 행렬 분해 기반 음향 신호 분리 기술 연구. 『전자공학회지』, 43(5), 25-34.
http://www.dbpia.co.kr/DownloadSociety/Download/NODE06681648

Multi-channel BSS on tensor representation

› Independent Component Analysis



› Independent Vector Analysis

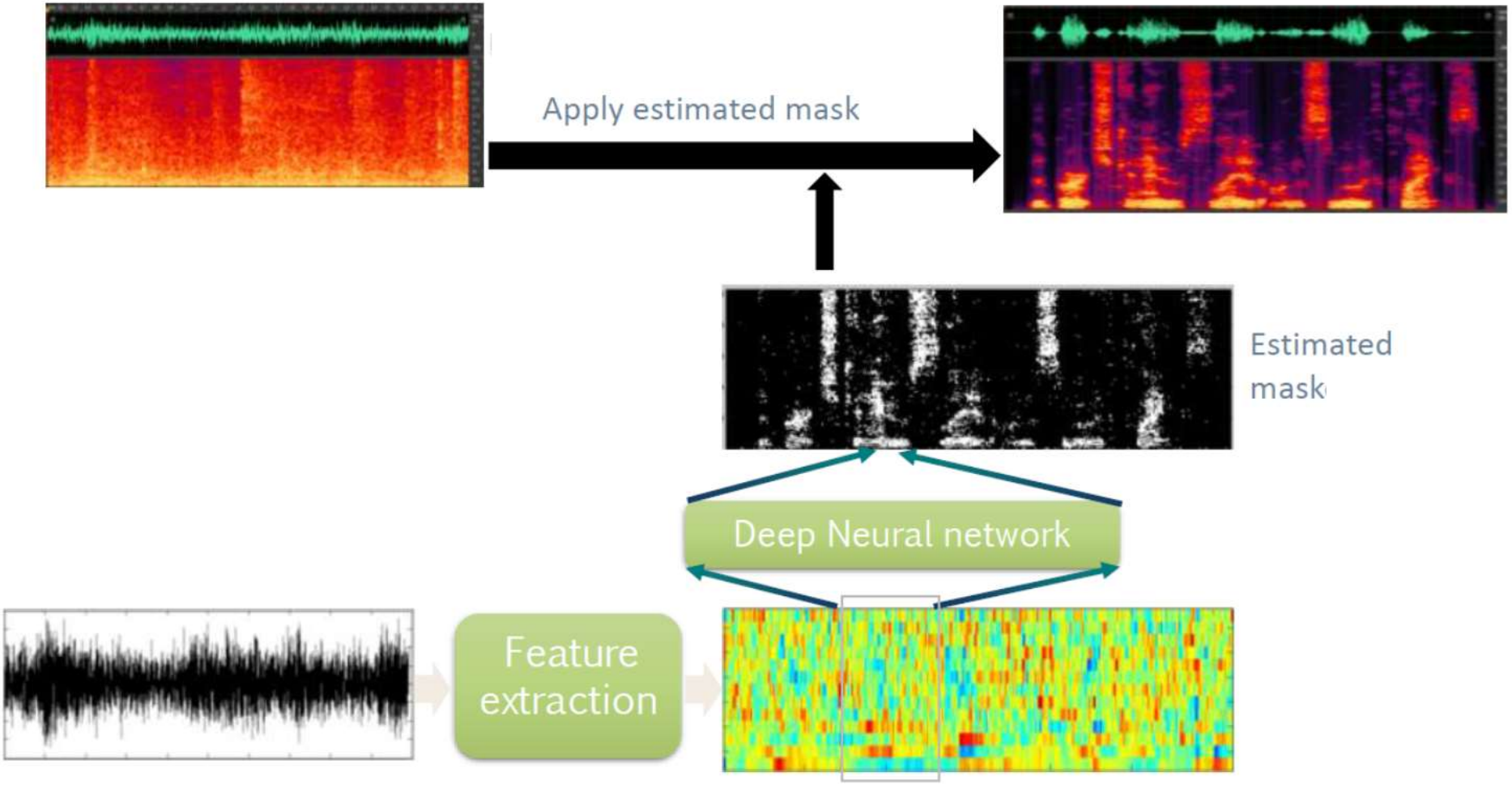


https://youtu.be/dQwTn98TuA?t=1m14s
KBS 스페셜: 제5의도시 중

H. Sawada et al, Blind Audio Source Separation on Tensor Representation, Tutorial ICASSP 2018
T. Kim, Independent Vector Analysis, Ph.D. Thesis, KAIST 2006

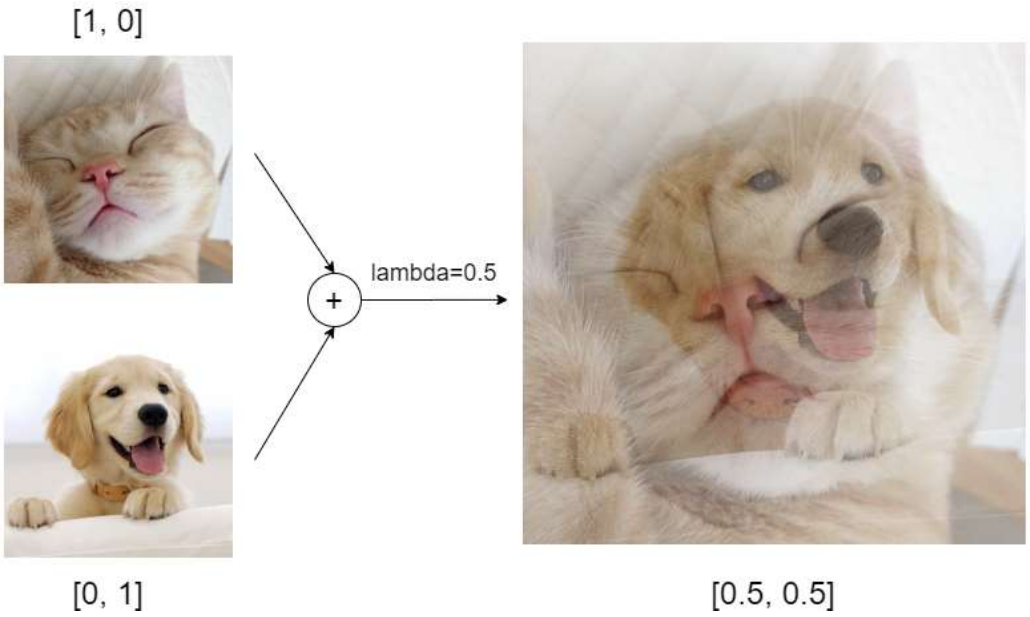
# Audio signal enhancement

- T. Kim (Neosapience), “Introduction to Speech Separation”
  - https://www.youtube.com/watch?v=OgNSFKeHy8k



# Comparison to Image enhancement..???

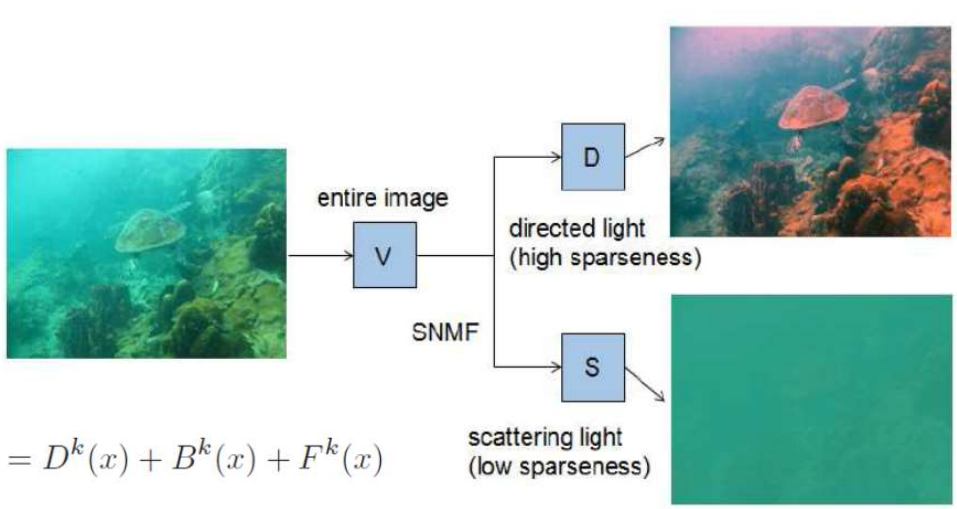
- H. Zhang, '**mixup**: Beyond Empirical Risk Minimization'
- https://medium.com/@wolframalphav1.0/easy-way-to-improve-image-classifier-performance-part-1-mixup-augmentation-with-codes-33288db92de5

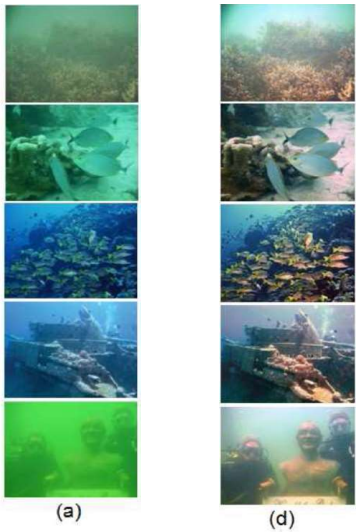




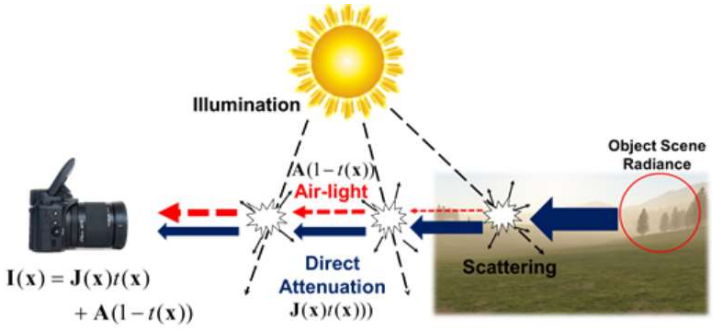
# Comparison to Image enhancement

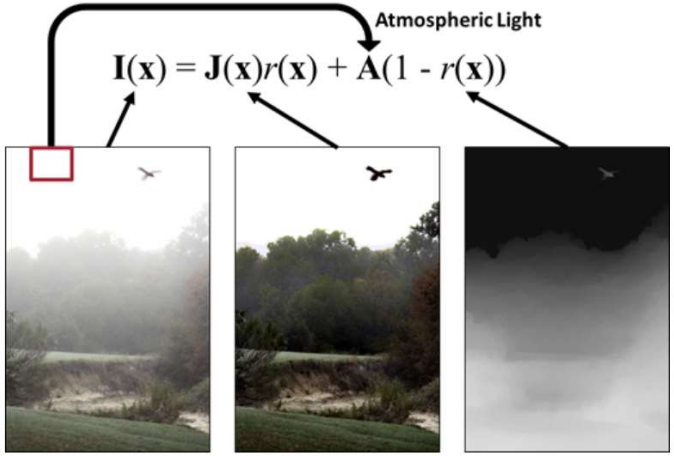
- X. Liu, ‘A Novel Underwater De-scattering Method Based on Sparse Non-negative Matrix Factorization’




- D. Park, ‘Single image haze removal with WLS-based edge-preserving smoothing filter’



# Y. Xu, ‘A Regression Approach to Speech Enhancement Based on Deep Neural Networks’

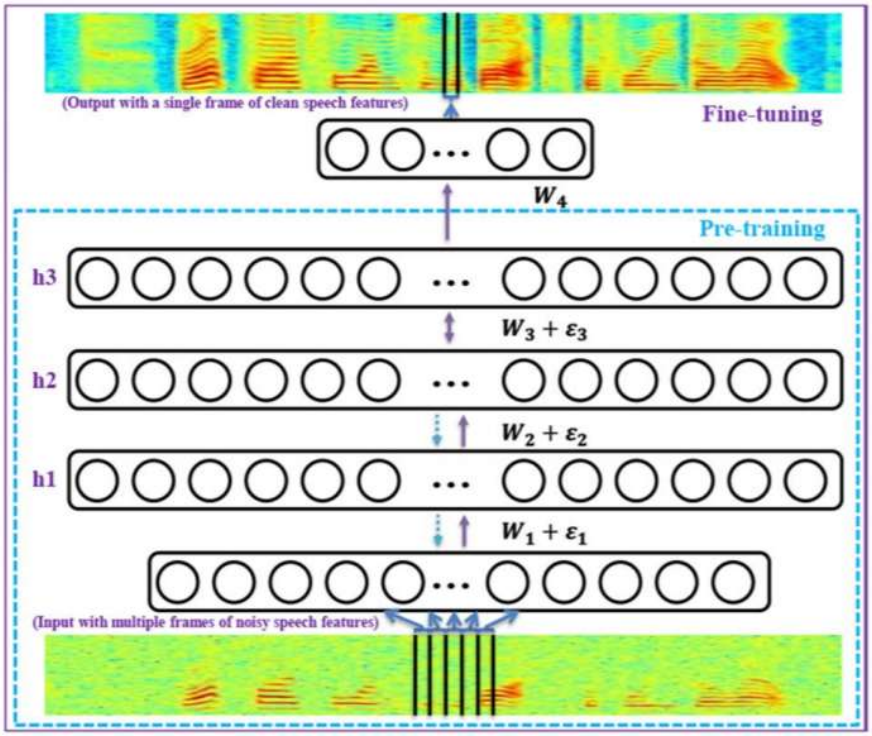
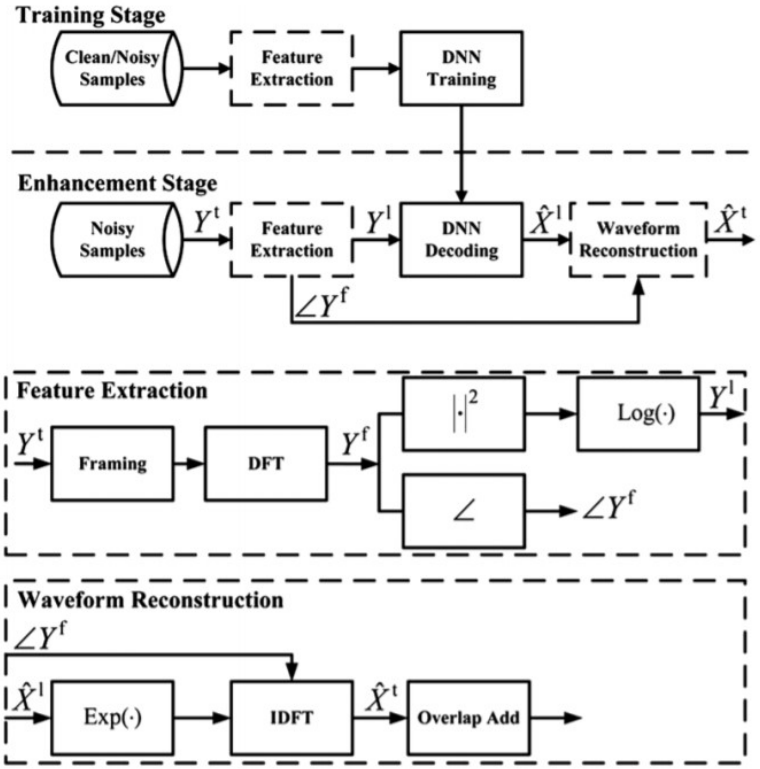
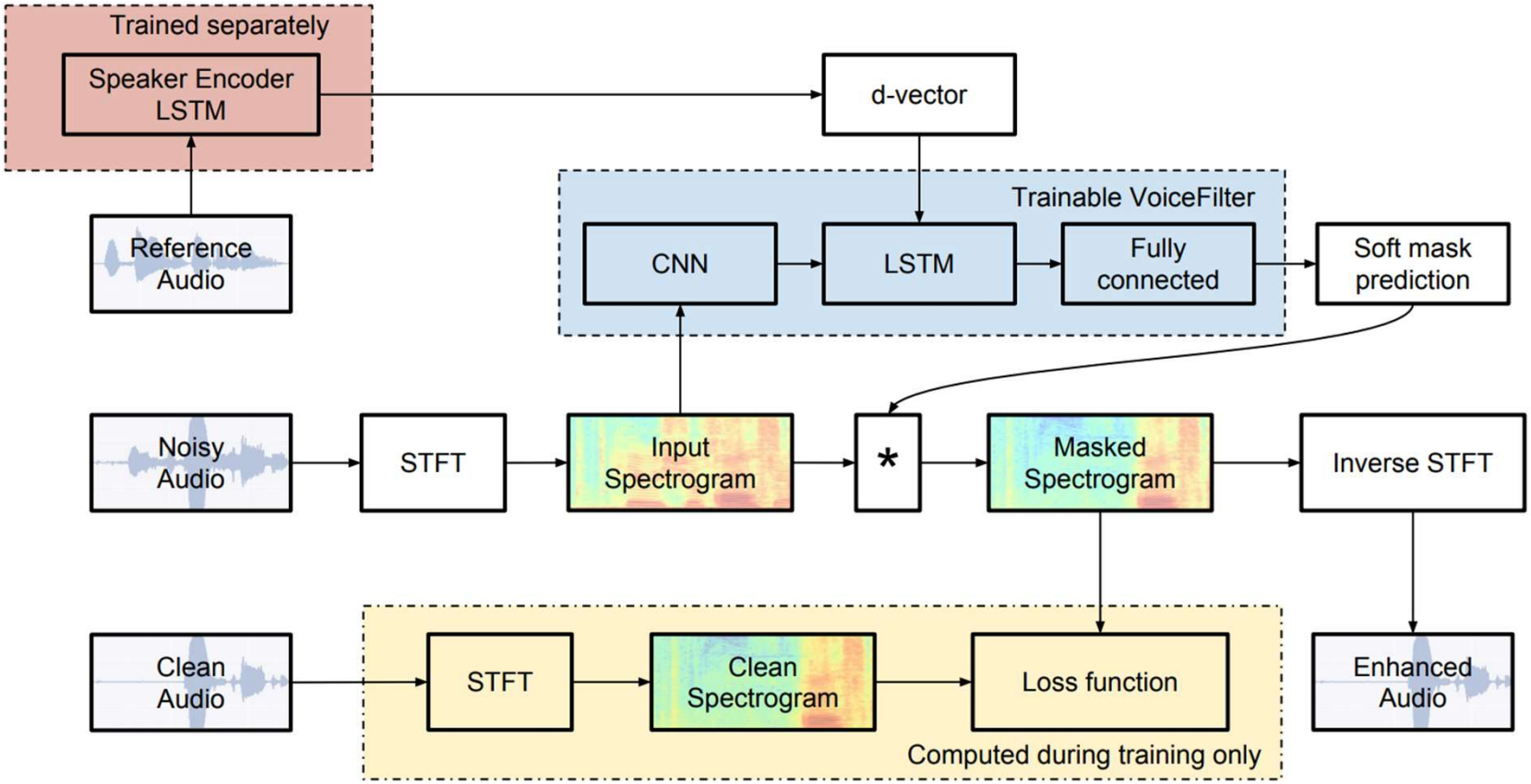



Fig. 2. Illustration of the basic DNN training procedure.



# Q.Wang, 'VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking'

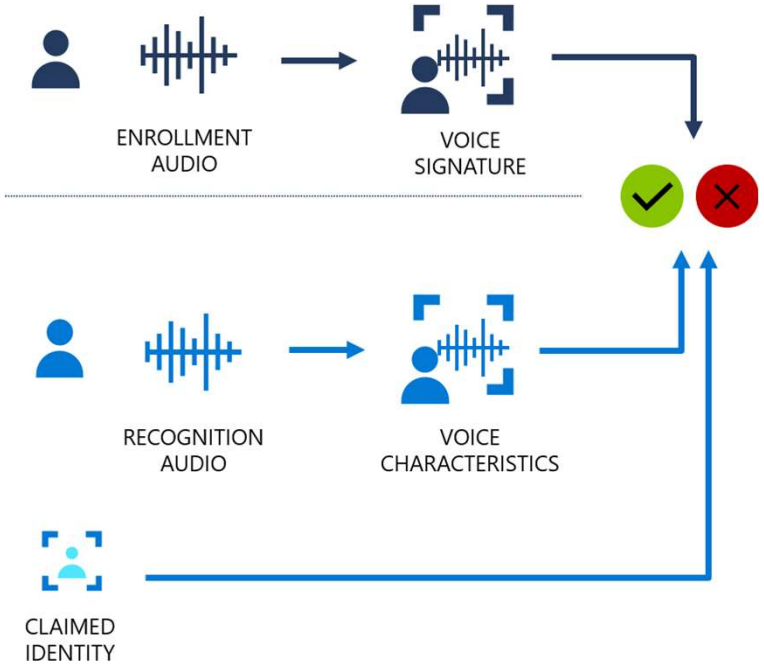
- Speaker recognition task을 위해 훈련된 network embedding vector
- Speech enhancement task를 위해서 입력으로 사용




- https://google.github.io/speaker-id/publications/VoiceFilter/
- https://www.youtube.com/watch?v=BiWMZdnHuVs&feature=emb_logo

# Speaker recognition

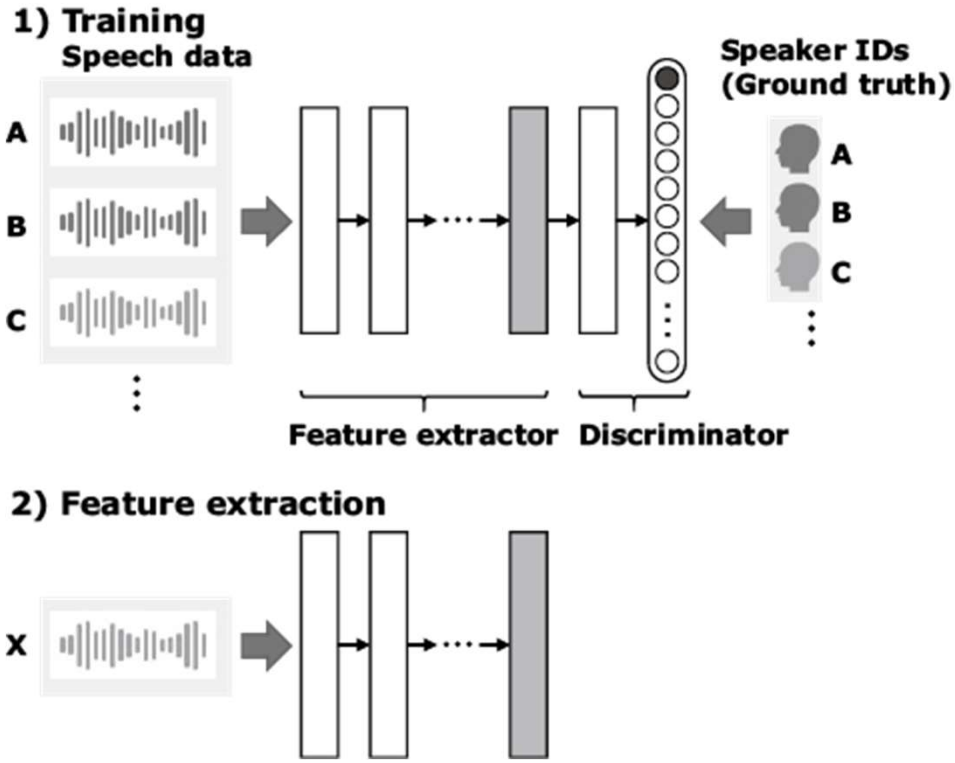


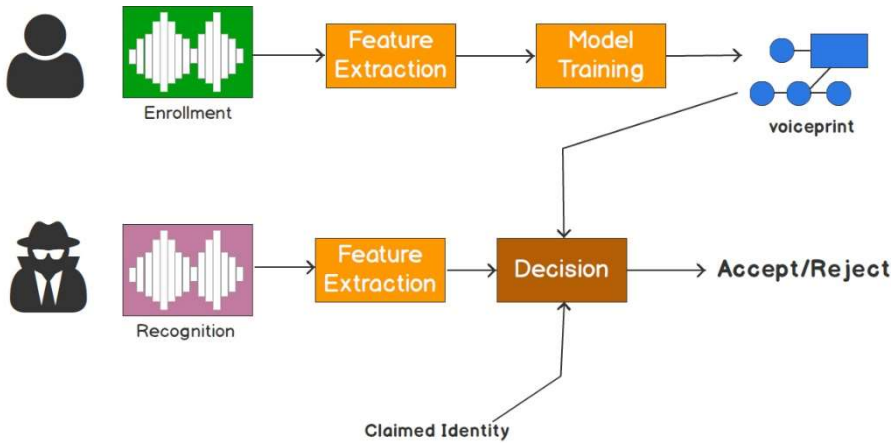



- https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/speaker-recognition-overview
- https://wiki.aalto.fi/display/ITSP/Speaker+Recognition+and+Verification

# Speaker recognition

Metric learning… [*]




- https://www.nec.com/en/global/techrep/journal/g18/n02/180218.html
- https://www.mathworks.com/matlabcentral/fileexchange/69617-speaker-recognition-biometric-system-matlab-code
- [*] J. Chung, “In defence of metric learning for speaker recognition"

# 특정 **사람의 목소리**만 추출 (enhancement)

1. Pre-trained **speaker** recognition network 준비
2. **특정 화자**의 network embedding을 뽑는다
3. Noisy input 과 embedding을 네트워크에 넣는다
4. Enhanced output을 얻는다

< RIP… >

▶ Generalized Gamma distribution [J. Shin, 2005]

$$p(X_k) = \frac{\gamma^2 \beta^{2\eta}}{4\Gamma(\eta)^2} |X_{k,R} X_{k,I}|^{\eta\gamma-1} \cdot \exp(-\beta|X_{k,R}|^\gamma - \beta|X_{k,I}|^\gamma)$$

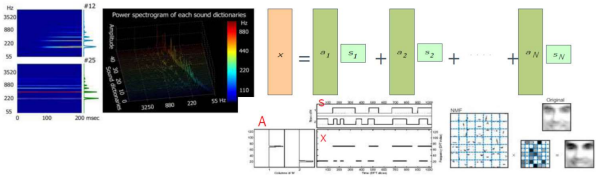
$$\Lambda_k = \frac{p(X_k|H_1)}{p(X_k|H_0)} = \frac{\hat{\gamma}_S^2 \hat{\beta}_S^{2\hat{\eta}_S} \Gamma(\hat{\eta}_N)^2}{\hat{\gamma}_N^2 \hat{\beta}_N^{2\hat{\eta}_N} \Gamma(\hat{\eta}_S)^2} |X_R X_I|^{\hat{\eta}_S\hat{\gamma}_S - \hat{\eta}_N\hat{\gamma}_N} \cdot e^{(-\hat{\beta}_S(|X_R|^{\hat{\gamma}_S}+|X_I|^{\hat{\gamma}_S}) + \hat{\beta}_N(|X_R|^{\hat{\gamma}_N}+|X_I|^{\hat{\gamma}_N}))}$$

▶ Convex Combination of Multiple Statistical Models with Application to VAD [T. Petsatodis, 2011]

$$\Lambda_k^{\text{Convex}} \equiv w_G \Lambda_k^G + w_L \Lambda_k^L + w_\Gamma \Lambda_k^\Gamma = w_G \frac{f^{(G)}_{X_k|H_1}(X_k)}{f_{X|H_0}(X_k)} + w_L \frac{f^{(L)}_{X_k|H_1}(X_k)}{f_{X|H_0}(X_k)} + w_\Gamma \frac{f^{(\Gamma)}_{X_k|H_1}(X_k)}{f_{X|H_0}(X_k)} = \frac{w_G f^{(G)}_{X_k|H_1}(X_k) + w_L f^{(L)}_{X_k|H_1}(X_k) + w_\Gamma f^{(\Gamma)}_{X_k|H_1}(X_k)}{f_{X|H_0}(X_k)}$$

Single channel approach in spectrogram

› Non-negative Matrix Factorization

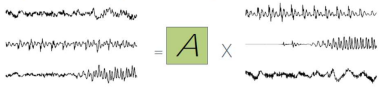


› Computational Auditory Scene Analysis
– Segmentation of auditory elements in spectrogram

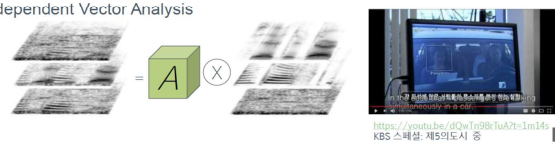
For more information about NMF approach in Korean
권기수, 강남수 (2016.5). 템플릿 기반 음향 신호 분리 기술 연구. 전자공학회지, 43(5), 25-34.
http://www.dbpia.co.kr/Download/Society/Download/NOD00568164B

Multi-channel BSS on tensor representation

› Independent Component Analysis



› Independent Vector Analysis



https://youtu.be/dQwTn98rTuA?t=1m14s
KBS 스페셜: 제5의도시 중

H. Sawada et al, Blind Audio Source Separation on Tensor Representation, Tutorial ICASSP 2018
T. Kim, Independent Vector Analysis, Ph.D. Thesis, KAIST 2006

# 특정 **소리**만 추출 (enhancement)

1. Pre-trained **sound recognition** network 준비
2. **특정 소리**의 network embedding을 뽑는다
3. Noisy input 과 embedding을 네트워크에 넣는다
4. Enhanced output을 얻는다

**ONE-SHOT CONDITIONAL AUDIO FILTERING OF ARBITRARY SOUNDS**

*Beat Gfeller, Dominik Roblek and Marco Tagliasacchi*

Google Research

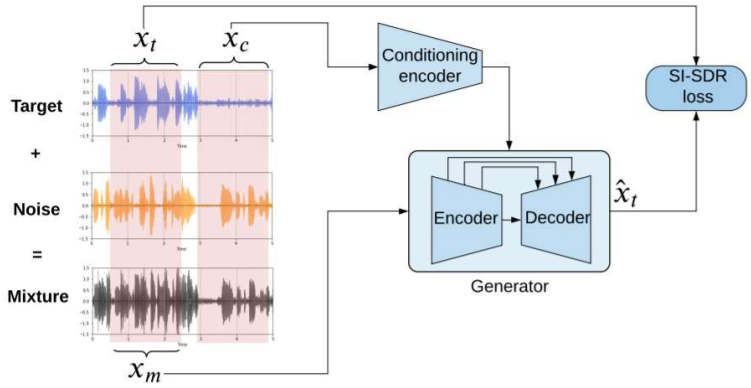
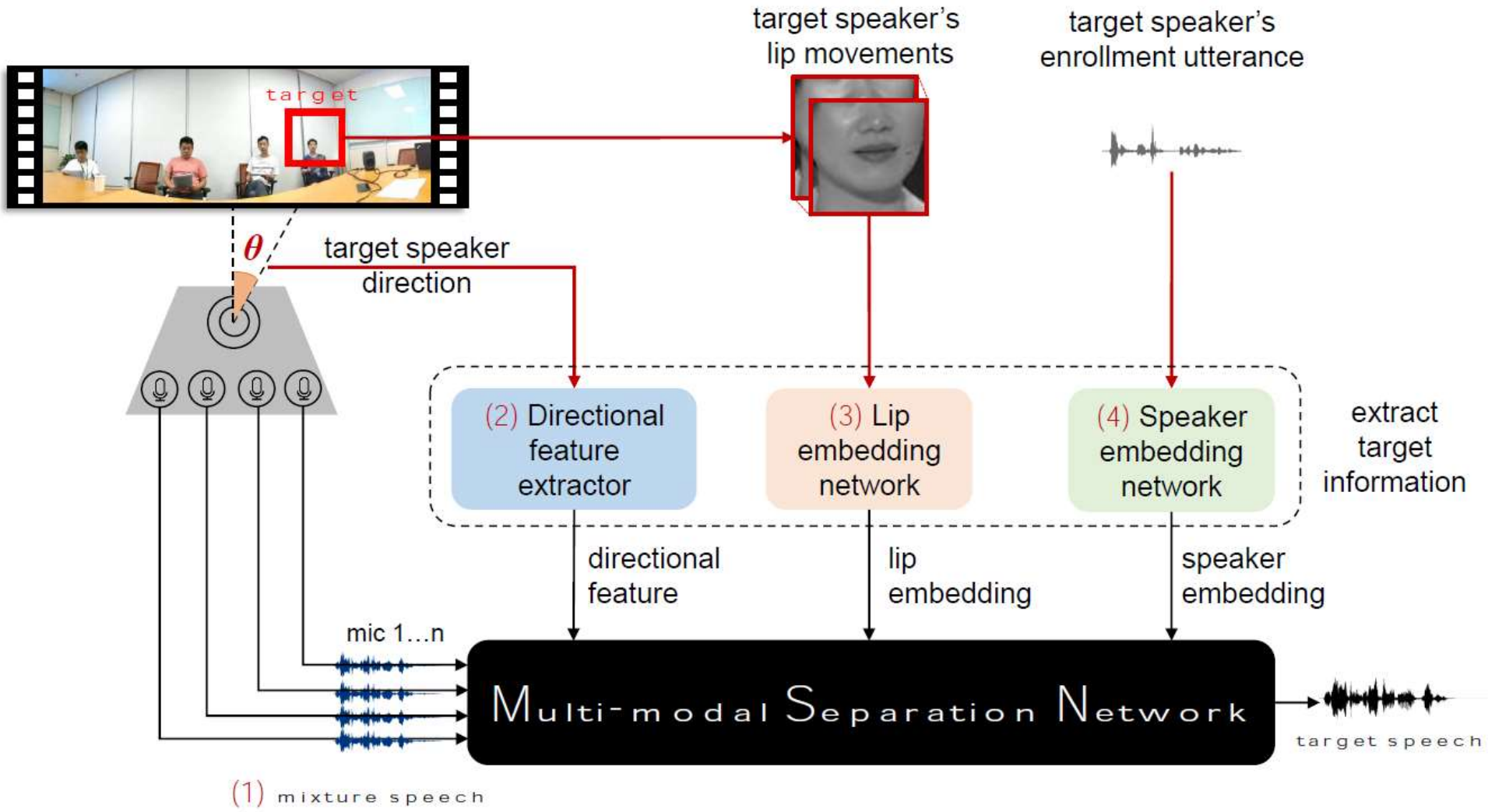


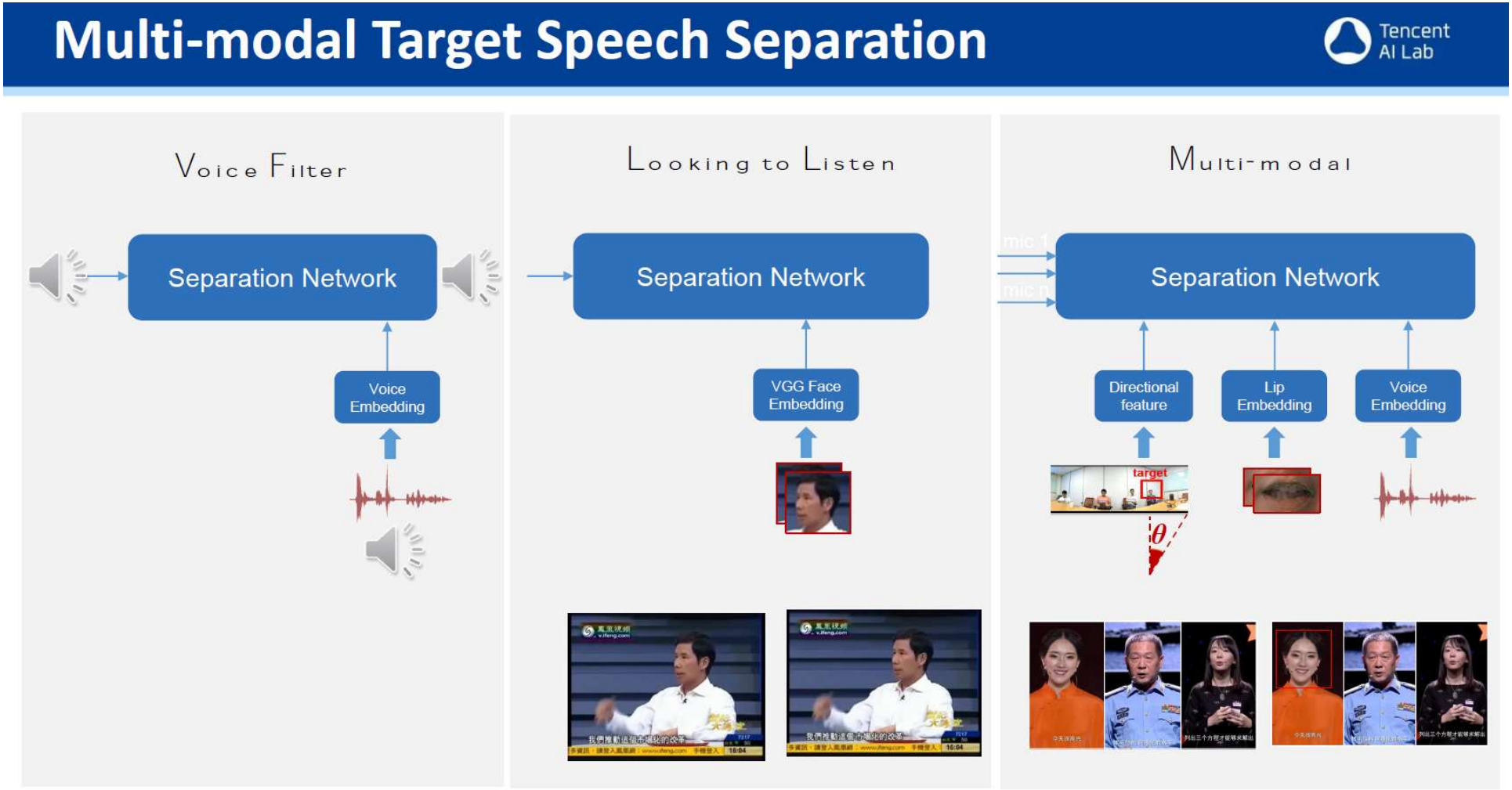

**Fig. 1**: *SoundFilter* model overview.

# Put any embedding!

- D, Yu (Tencent), “Solving Cocktail Party Problem” CHiME 6, 2020



# Put any embedding!



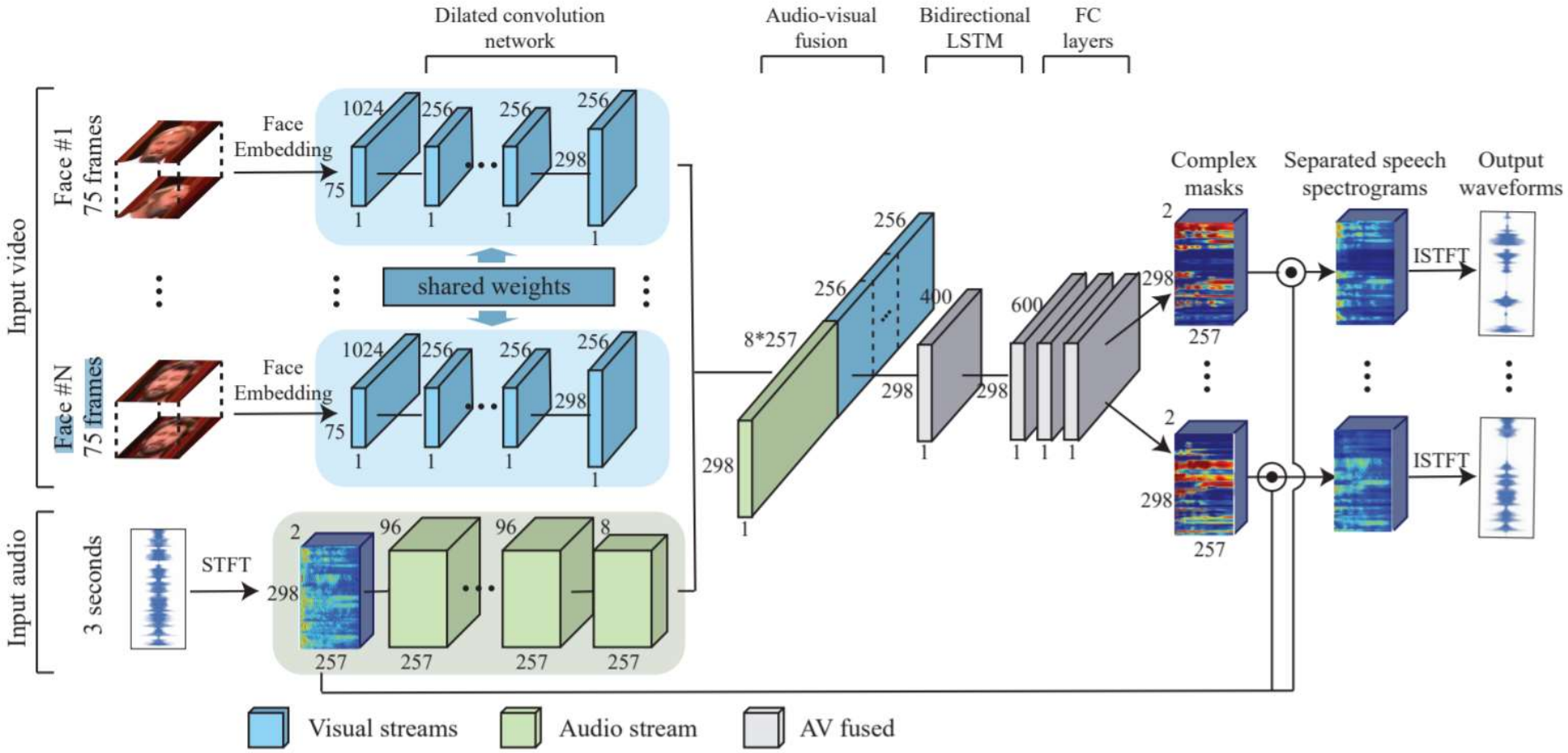





- https://jupiterethan.github.io/av-enh.github.io/

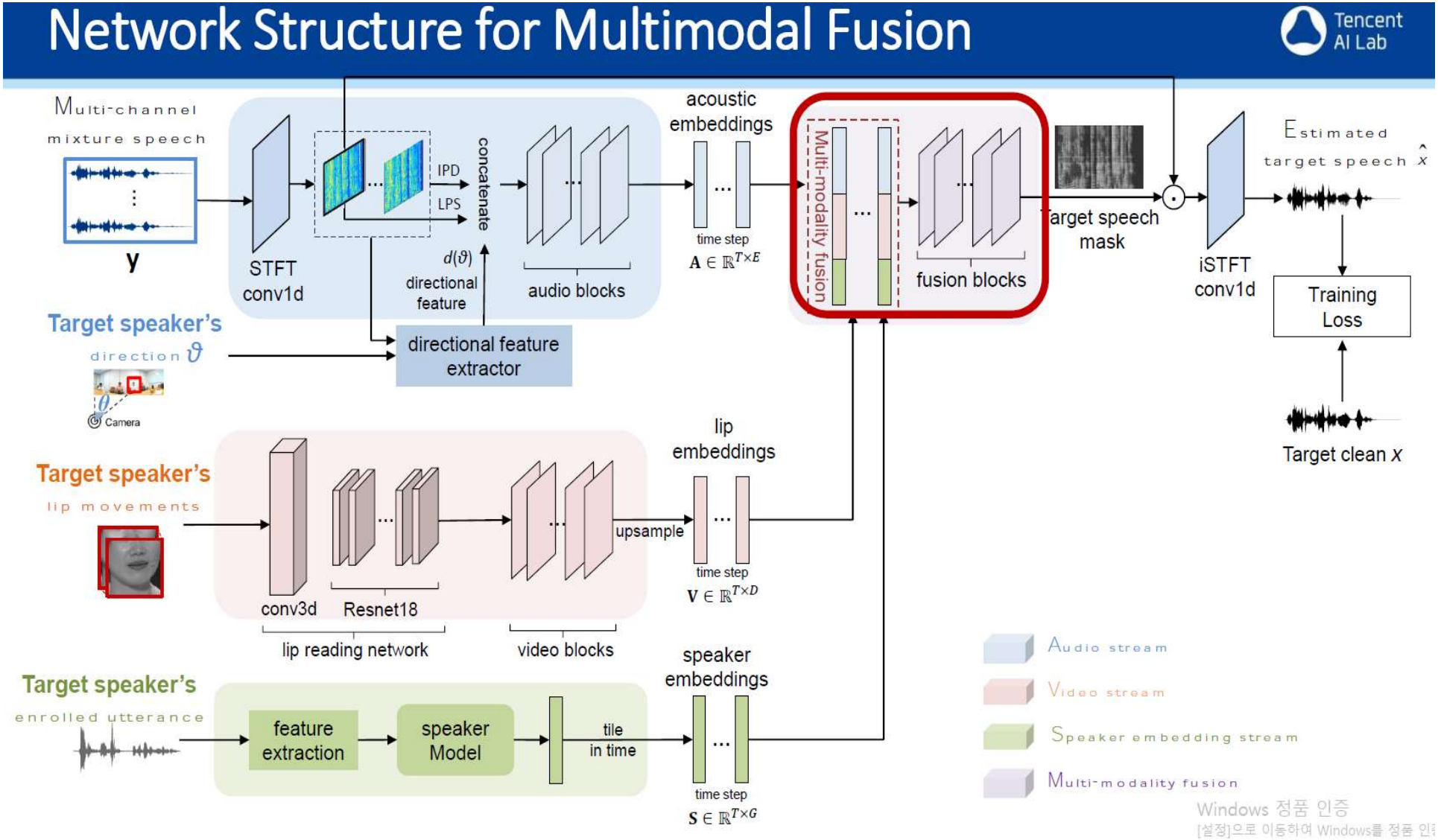
# Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation

ARIEL EPHRAT, Google Research and The Hebrew University of Jerusalem, Israel
INBAR MOSSERI, Google Research




- https://looking-to-listen.github.io/

# D, Yu (Tencent), “Solving Cocktail Party Problem” CHiME 6, 2020



# D, Yu (Tencent), “Solving Cocktail Party Problem” CHiME 6, 2020

(SDR 0.44 / PESQ 1.87)
Before enhancement : WER 48.9 [%]

| Input Modalities | | | SDR (dB) | PESQ | WER (%) |
|---|---|---|---|---|---|
| Directional Feature | Enrolled Voice | Target Lip | | | |
| √ | | | 16.9 | 3.24 | 11.3 |
| | √ | | 14.8 | 2.98 | 14.7 |
| | | √ | 16.6 | 3.01 | 19.6 |
| √ | √ | | 17.1 | 3.25 | 10.5 |
| √ | √ | √ | **17.6** | 3.28 | 10.0 |
| √ | | √ | **17.5** | 3.28 | 10.3 |

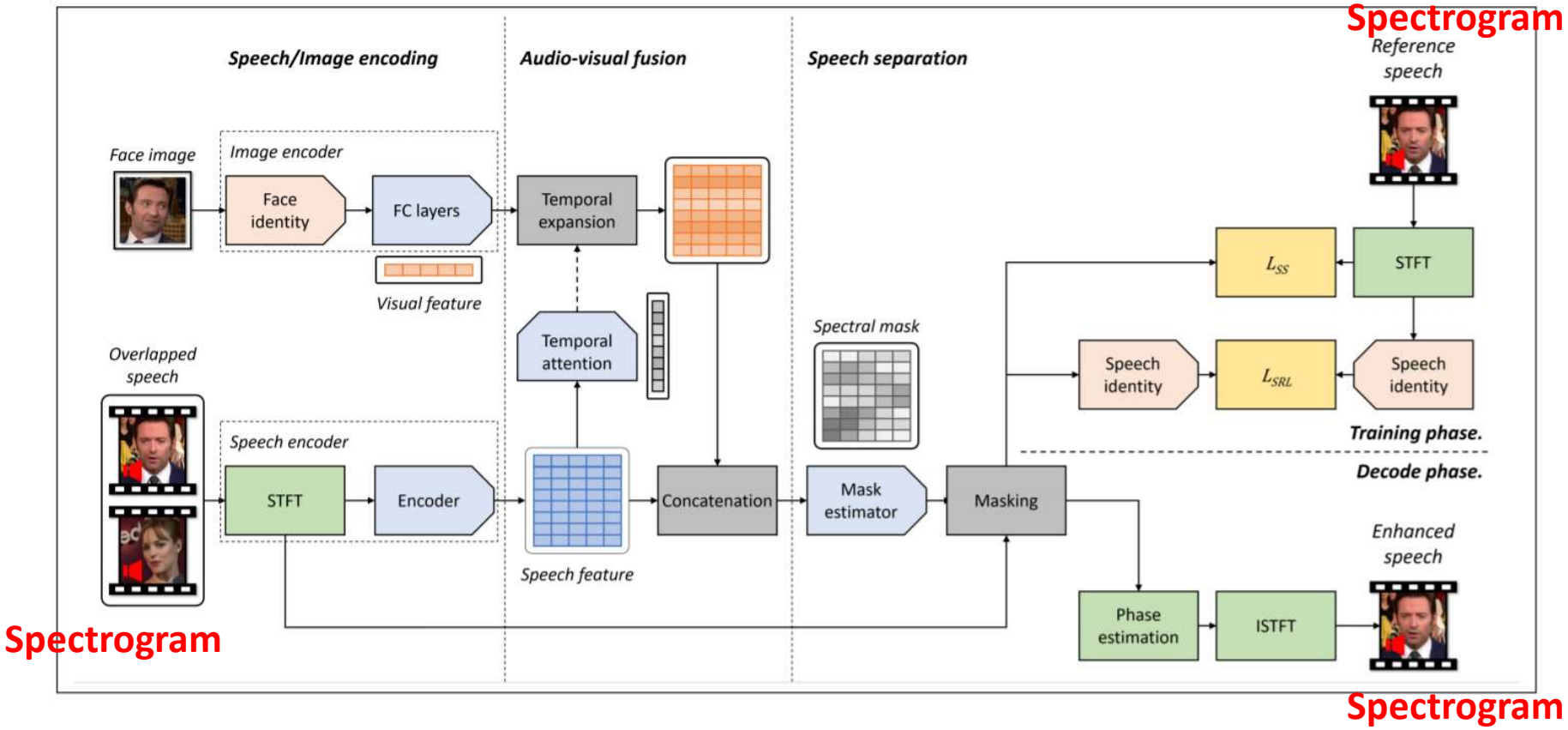
Difference from **speaker embedding** is small。

We can stick to use only Directional feature + Lip if speaker embedding not available

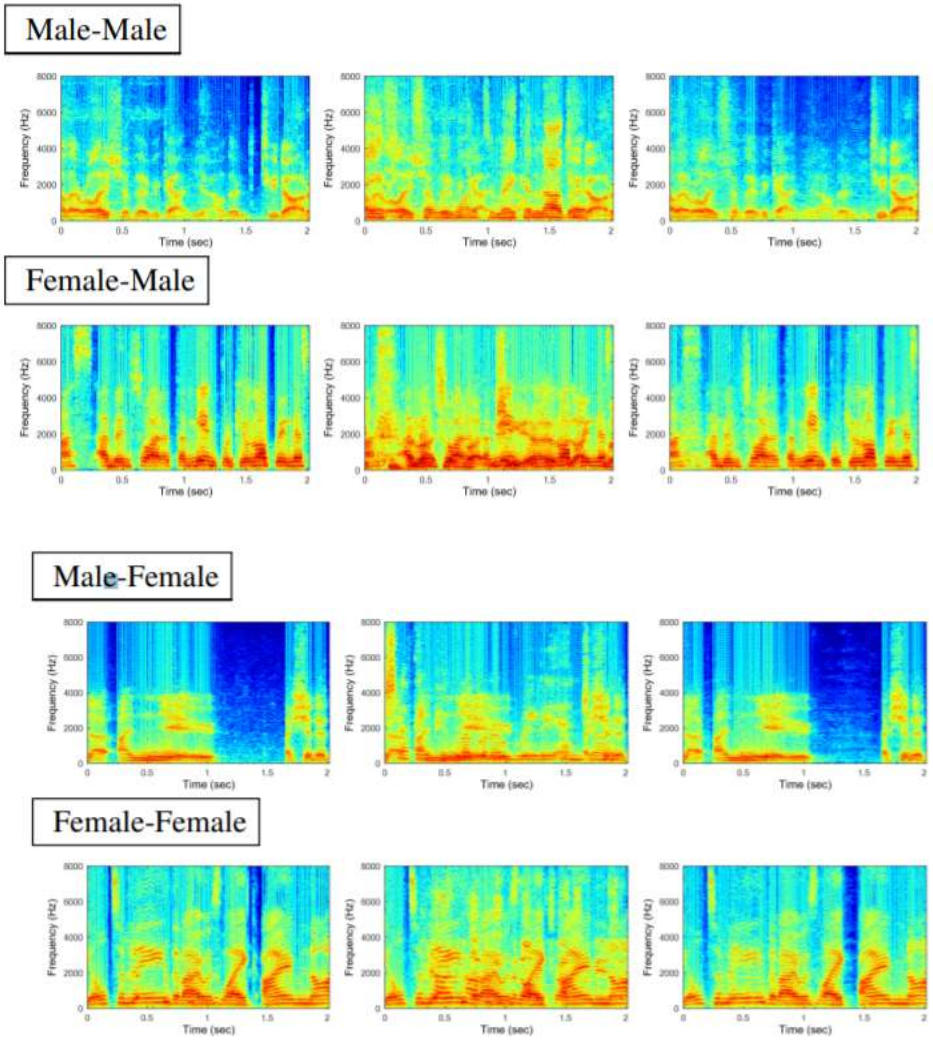
# S.Chung ‘FaceFilter: Audio-visual speech separation using still images’

- Sync가 확보된 동영상 입력에 대해서만 쓸 수 있나요?

→ 아뇨. 그냥 사전에 찍어 놓은 single image만으로, 목소리 분리가 가능합니다.



# S.Chung ‘FaceFilter: Audio-visual speech separation using still images’




| Mixture | SDRi |
|---|---|
| **Seen-Heard Speaker** | |
| Male-Male | 2.084 dB |
| Male-Female | 3.039 dB |
| Female-Male | 3.821 dB |
| Female-Female | 1.930 dB |
| **Unseen-Unheard Speaker** | |
| Male-Male | 1.256 dB |
| Male-Female | 3.066 dB |
| Female-Male | 3.830 dB |
| Female-Female | 1.198 dB |

# Summary

- Speech enhancement는 spectrogram (Time-Freq block)을 [0-1] 사이의 값으로 masking 하는 네트워크를 사용한다. (e.g. sigmoid output [Time x Freq] )

- 기본적인 신호처리 / 모델링을 넘어서,
  특정 목적에 부합하는 다양한 정보들의 embedding을 뽑고, 입력으로 넣어주자
  - 사전 녹음 된 특정 목소리 찾기 → VoiceFilter (Google, 2019)
  - 사전 녹음 된 특정 소리 찾기 → SoundFilter (Google, 2020)
  - 특정 화자의 영상에서 얼굴/입술 정보 → Looking to Listen (Google, 2018)
  - 특정 화자의 영상에서 얼굴/입술/방향/(사전)목소리 정보 → Audio-Visual Speech Separation and Dereverberation with a Two-Stage Multimodal Network (Tencent, 2020)
  - 사전 촬영 된 특정 얼굴의 목소리 찾기 → FaceFilter (Yonsei, 2020)

- ~~All you need is embedding~~
  - ~~특정 화자의 얼굴/입술/홍채/지문 embedding을 이용한 목소리 분리…~~