Time-Frequency Masking-based Speech Enhancement using Generative Adversarial Network

Meet H. Soni, Neil Shah, and Hemant A. Patil

ICASSP 2018, Paper id- 3043

Presented By-

Prof. Hemant A. Patil

Speech Research Lab,
Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT),
Gandhinagar, Gujarat, India.

20 April 2018

Presentation Outline

- Time-Frequency (T-F) masking for speech enhancement (SE)
- Task-dependent masking for SE with supervised learning methods
- Generative Adversarial Networks (GANs)
- Adversarial training (GAN) *vs.* maximum likelihood (ML)-based optimization
- T-F masking using vanilla GAN (v-GAN)
- T-F masking using proposed Minimum Mean Square Error GAN (MMSE-GAN)
- Performance comparisons with other approaches
- Summary and conclusions

Speech Enhancement - Motivation

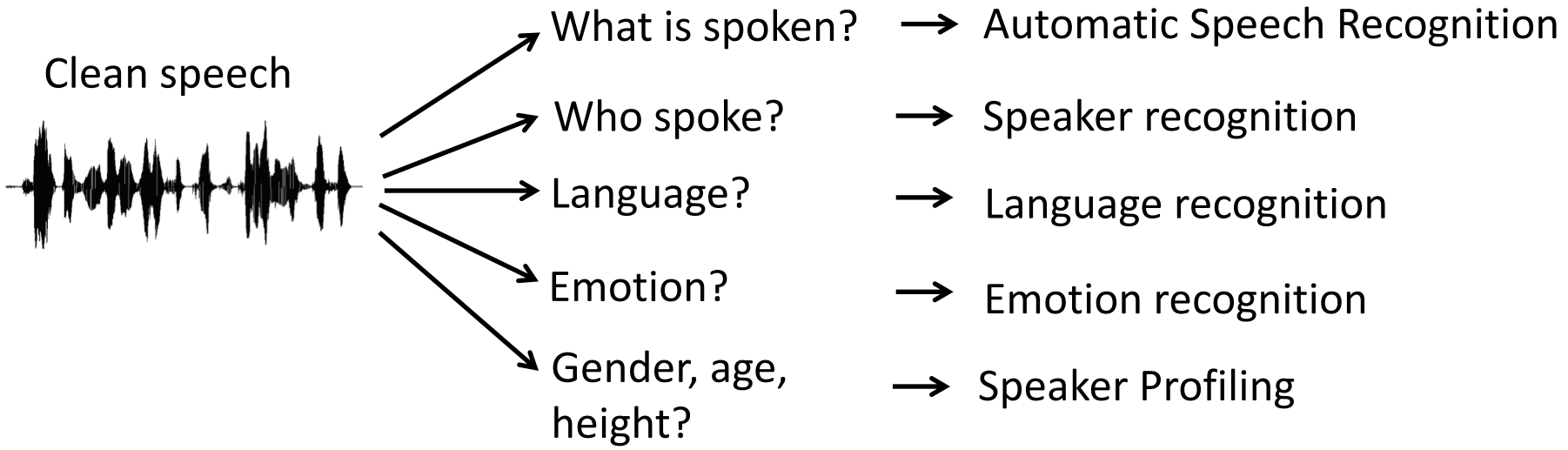


Fig 1. Various Speech Technology Applications.

- What if the speech is noisy?
  - Would the message be clear?
  - Would the speaker be recognized properly?
  - What if we enhance the clean speech components present in the noisy mixture?

Source:

Weninger, Felix, et al. ”Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR”, International Conference on Latent Variable Analysis and Signal Separation, Springer, Cham, 2015, pp. 91-99.

Speech Enhancement contd.

Speech Enhancement (SE)

- Enhancing a real-life degraded speech
- Improving the speech intelligibility and quality given the noisy speech

Two kinds of enhancement techniques:

1. *Enhancement for machines*: for task, such as, ASR
2. *Enhancement for humans*: for task, such as, hearing aid design

Source:

A. A. Nugraha, A. Liutkus, and E. Vincent. ”Multichannel audio source separation with deep neural networks”, IEEE Trans. Audio, Speech, Lang. Process. (*TASLP*), vol. 24, no. 9, pp. 1652-1664, 2016.

Speech Enhancement - Literature

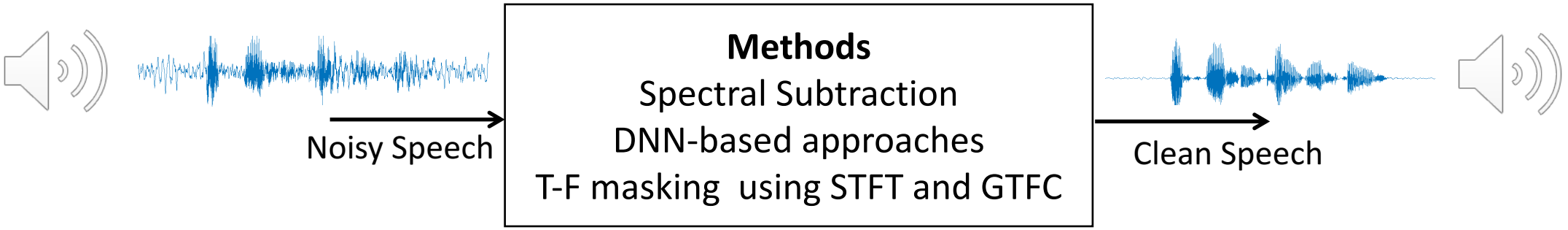


Fig 2. Trends in speech enhancement

Source:

- Yang, L. Ping, and Qian-Jie Fu.,"Spectral subtraction-based speech enhancement for cochlear implant patients in background noise", The Journal of the Acoustical Society of America (*JASA*), vol. 117, no. 3, pp. 1001-1004, 2005.
- A. A. Nugraha, A. Liutkus, and E. Vincent., "Multichannel audio source separation with deep neural networks", IEEE Trans. Audio, Speech, Lang. Process. *(TASLP)*, vol. 24, no. 9, pp. 1652-1664, 2016.
- B. Valentini, Cassia, et al., "Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech," in 9th ISCA Speech Synthesis Workshop *(SSW)*, Sep. 13-15, Sunnyvale, CA, USA, 2016.
- Y. Wang, A. Narayanan, and D. Wang., "On training targets for supervised speech separation", IEEE Trans. Audio, Speech, Lang. Process. *(TASLP)*, vol. 22, no. 12, pp. 1849-1858, Dec. 2014.

Speech Enhancement - Literature

Trends	Year	Comments
1. Spectral subtraction 2. Weiner filtering 3. Time-Frequency (TF) mask estimation using STFT, GTFC, etc.	1978-2010	- Do not improve speech intelligibility - Not suitable for highly stationary and low SNR noise conditions
T-F mask estimation using supervised learning techniques	2010-2016	- High computational cost - Better in separating background interferences - Not generalizable to unknown noisy conditions
Speech enhancement GAN (SEGAN) (a non-masking approach)	INTERSPEECH 2017	- Highly complex training involved - End-End speech enhancement (SE) technique in time domain - Does not follow mask estimation technique which is proven to perform better in SE tasks
Speech enhancement GAN (MMSE-GAN) **(Proposed masking approach)**	ICASSP 2018	- A simple T-F mask-based technique - A DNN-based GAN is employed - Have shown a need of incorporating MMSE regularizer in the vanilla GAN architecture - Estimates mask more accurately than the DNN-based techniques

Time-Frequency (T-F) mask

$$\begin{bmatrix} \text{Noisy T} - \text{F} \\ \text{Representation} \end{bmatrix} \odot \begin{bmatrix} \text{Mask} \end{bmatrix} = \begin{bmatrix} \text{Enhanced T} - \text{F} \\ \text{Representation} \end{bmatrix}$$

$\odot$: Element wise multiplication

- An ideal T-F mask modulates the T-F unit of the noisy spectrum
- Different types of training targets (T-F masks):
  1. Ideal Binary Mask (IBM)
  2. Target Binary Mask (TBM)
  3. **Ideal Ratio Mask (IRM)**
  4. Gammatone Frequency Power Spectrum (GF-POW)
  5. Short-Time Fourier Transform Spectral Magnitude (FFT-MAG) and Mask (FFT-MASK)

Source:

Y. Wang, A. Narayanan, and D. Wang., ”On training targets for supervised speech separation”, IEEE Trans. Audio, Speech, Lang. Process. (*TASLP*), vol. 22, no. 12, pp. 1849-1858, Dec. 2014.

Time-Frequency (T-F) mask

Ideal Ratio Mask (IRM)

$$IRM(t,f) = \left(\frac{S^2(t,f)}{(S^2(t,f) + N^2(t,f))} \right)^{\beta} ,$$

where $S^2(t,f)$ and $N^2(t,f)$ denote the speech and noise energy, respectively, in particular T-F unit.

β is a tunable parameter.

Ideally, $\beta = 0.5$ is the best choice.

With $\beta = 0.5$, the equation becomes similar to the square root Wiener filter.

Source:

Y. Wang, A. Narayanan, and D. Wang., ”On training targets for supervised speech separation”, IEEE Trans. Audio, Speech, Lang. Process. (*TASLP*), vol. 22, no. 12, pp. 1849-1858, Dec. 2014.

T-F masking generalized block diagram

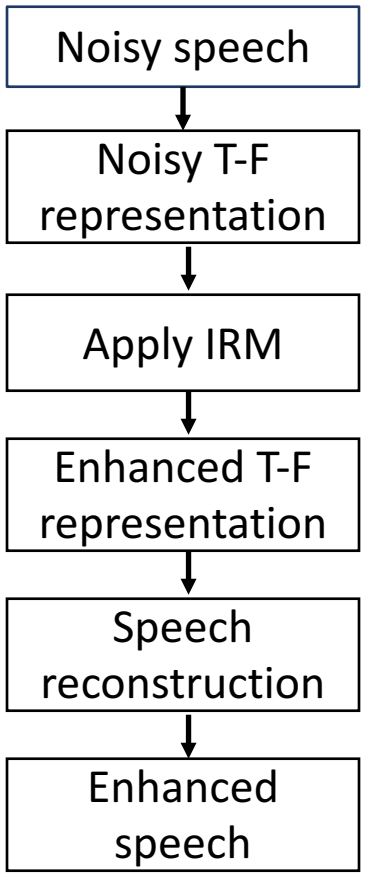


Fig 3. The general schematic of T-F masking approach.

Source:

Y. Wang, A. Narayanan, and D. Wang., ”On training targets for supervised speech separation”, IEEE Trans. Audio, Speech, Lang. Process. (*TASLP*), vol. 22, no. 12, pp. 1849-1858, Dec. 2014.

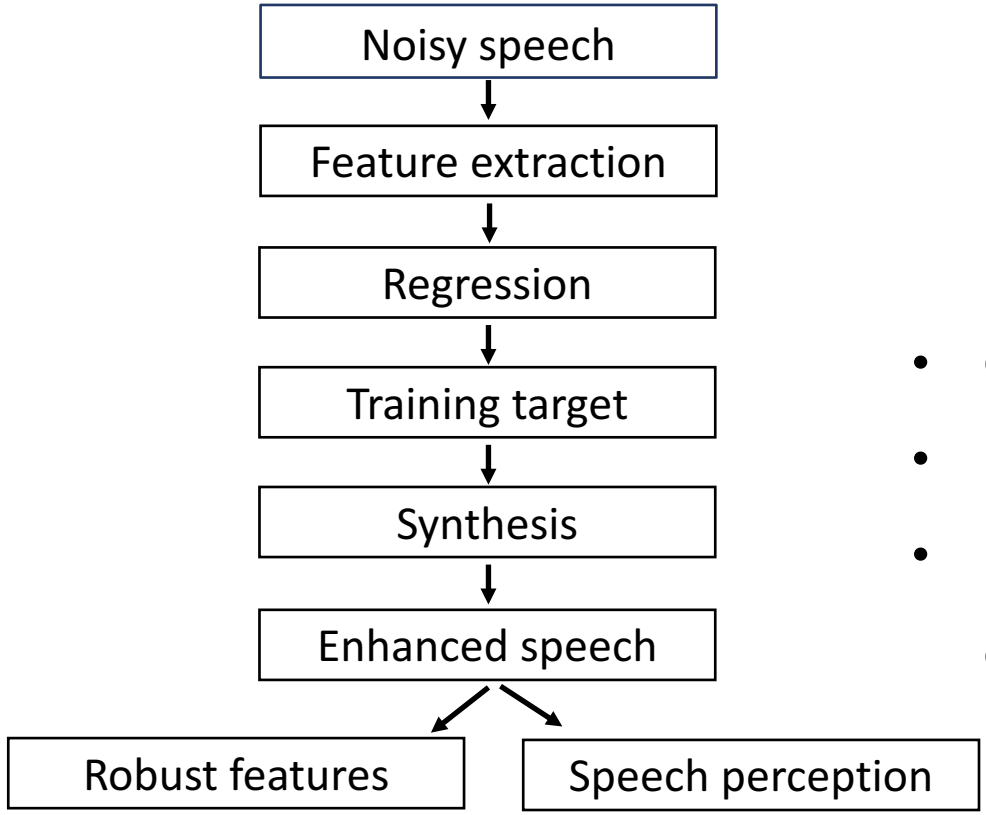
T-F masking approaches

1. T-F masking using Short Time Fourier Transform (STFT) and Gammatone Frequency Coefficients (GTFC)
    - Only works for additive noise
    - Mask cannot be calculated for the real-life noisy mixture, due to unavailability of clean signal
    - Considers only the magnitude information and ignores the phase information, leading to phase distortion
2. T-F masking using supervised learning algorithms

Source:

- Y. Wang, A. Narayanan, and D. Wang., ”On training targets for supervised speech separation”, IEEE Trans. Audio, Speech, Lang. Process. (*TASLP)*, vol. 22, no. 12, pp. 1849-1858, Dec. 2014.
- Meet H. Soni, Neil Shah, and Hemant A. Patil, “Time-Frequency masking-based speech enhancement using Generative Adversarial Network”, to appear in the, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Alberta, Canada, 2018.

T-F masking using supervised learning



- Generalizable to any T-F representation
- No need of noisy phase for reconstruction
- No need of clean speech in mask calculation once trained on few samples

Fig 4. The general block-diagram of using supervised learning in SE task.

Source:

- Y. Wang, A. Narayanan, and D. Wang., ”On training targets for supervised speech separation”, IEEE Trans. Audio, Speech, Lang. Process. *(TASLP)*, vol. 22, no. 12, pp. 1849-1858, Dec. 2014.
- Chen, Jitong, Yuxuan Wang, Sarah E. Yoho, DeLiang Wang, and Eric W. Healy., ”Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises”, The Journal of the Acoustical Society of America (*JASA*). vol. 139, no. 5, pp. 2604-2612, 2016.

T-F masking using supervised learning (before 2014)

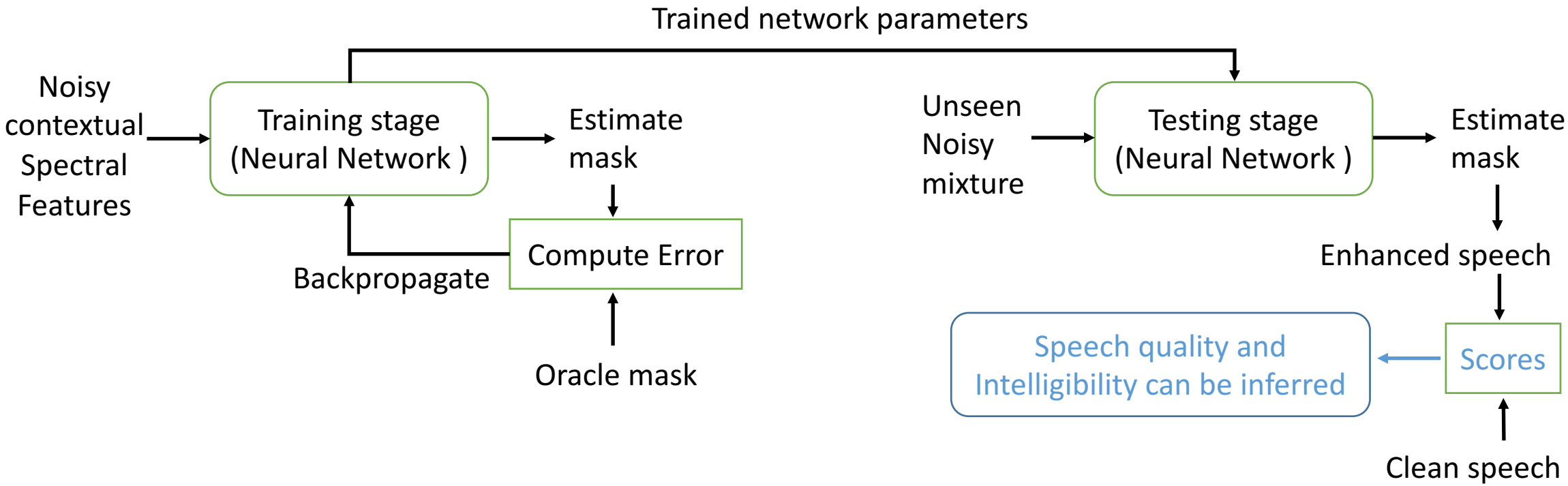


Fig 5. Training and testing procedure employed in SE task.

Source:

Y. Wang, A. Narayanan, and D. Wang., "On training targets for supervised speech separation", IEEE Trans. Audio, Speech, Lang. Process. (*TASLP*), vol. 22, no. 12, pp. 1849-1858, Dec. 2014.

T-F masking using supervised learning – *Task-dependent masking* technique in 2014

- Do not train the network with respect to the oracle mask

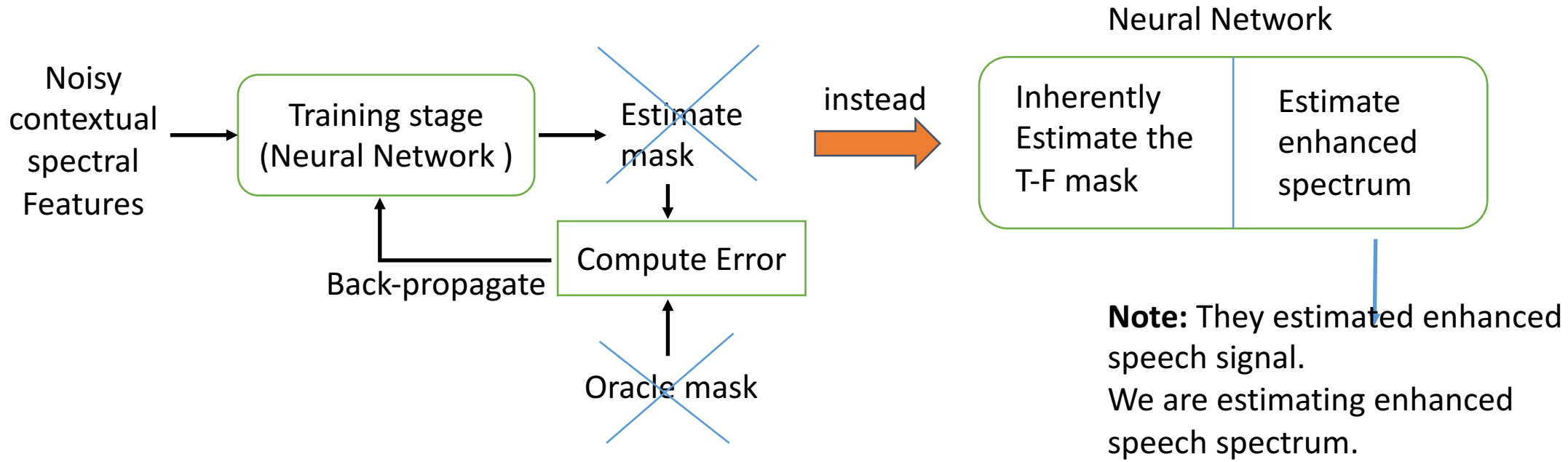
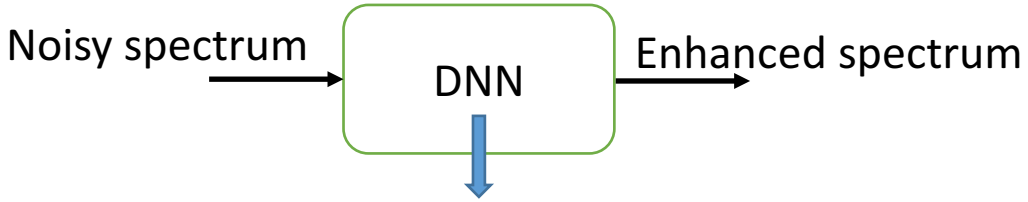


Fig 6. Inherent estimation of the T-F mask.

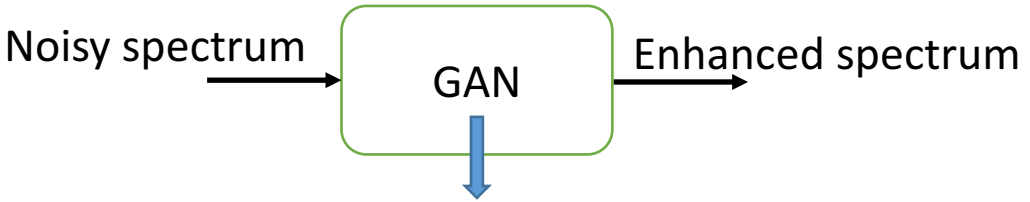
Source:

Wang, Yuxuan, and D. Wang., "A Neural Network For Time-Domain Signal Reconstruction: Towards Improving The Perceptual Quality Of Supervised Speech Separation." Department of Computer Science and Engineering, The Ohio State University, Tech. Rep (2014).

GAN *vs.* ML-based optimization networks



- Maximum likelihood (ML)-based optimization
- MMSE assumes output variables to be Gaussian
- This assumption prevents the network from learning perceptually optimum parameters
- Massive difference between oracle and estimated mask



- A generative modeling technique
- Alternative to ML-based optimization criteria.
- Learns perceptually optimum parameters
- Predicts an unknown distribution at the input
- Proven to show improvement over DNN, in voice conversion and speech enhancement tasks

Source:

- T. Kaneko, et al., ”Sequence-to-Sequence Voice Conversion with Similarity Metric Learned Using Generative Adversarial Networks”, in Proc. of the Int. Speech Communication Association Conf. (INTERSPEECH), Stockholm, Sweden, 2017, pp. 1283-1287.
- Hsu, Chin-Cheng, et al., ”Voice Conversion from Unaligned Corpora using Variational Autoencoding Wasserstein Generative Adversarial Networks”, in Proc. of the INTERSPEECH, Stockholm, Sweden, 2017, pp. 3364-3368.

Generative Adversarial Network (GAN)

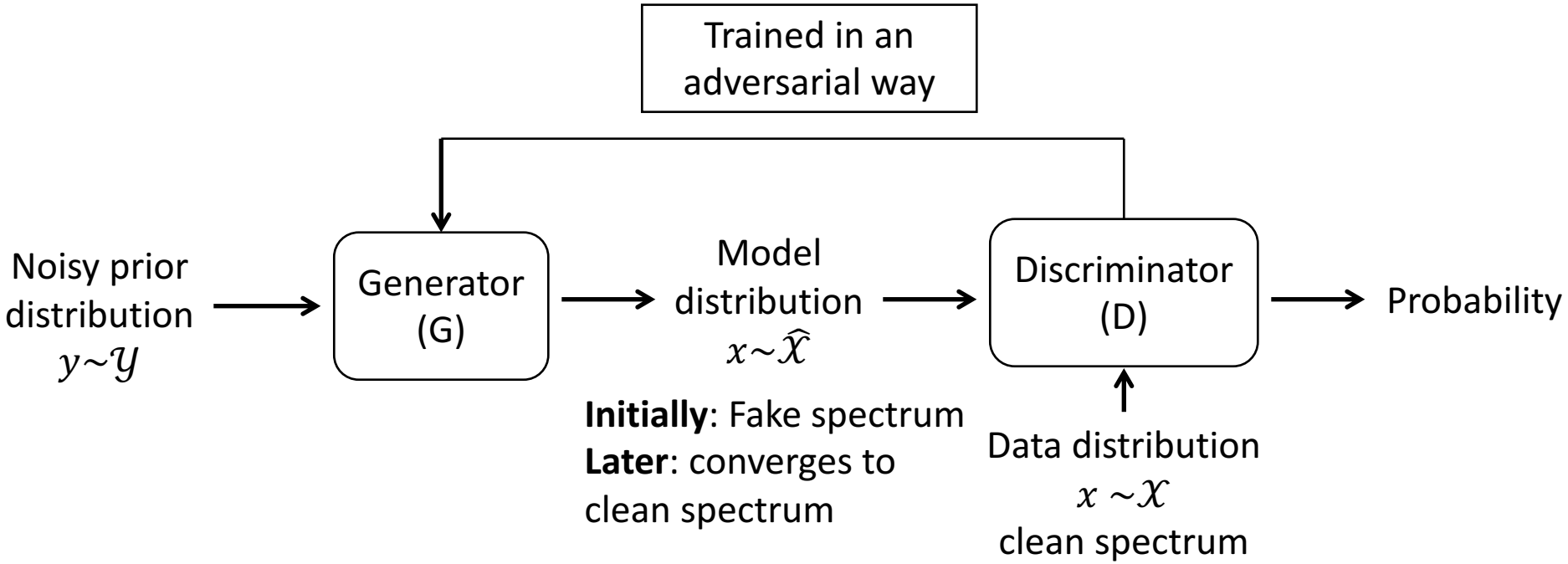


Fig 7. The general schematic of GAN exploited in SE task.

$$\min_G V(G) = -\mathbb{E}_{y\sim\mathcal{Y}}\left[\log D\big(G(y)\big)\right] \dots\dots\dots (1)$$

$$\min_D V(D) = -\mathbb{E}_{x\sim\mathcal{X}}[\log D(x)] - \mathbb{E}_{y\sim\mathcal{Y}}\left[1 - \log D\big(G(y)\big)\right] \dots\dots\dots\dots (2)$$

where $\mathbb{E}_{y\sim\mathcal{Y}}$: Expectation over all the samples y coming from distribution $\mathcal{Y}$

Source:

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Bengio, Y, "Generative adversarial nets", in Advances in Neural Information Processing Systems *(NIPS)*, 2014, pp. 2672-2680.

Generative Adversarial Network (GAN)

$$\min_G V(G) = -\mathbb{E}_{y\sim y}\left[\log D\big(G(y)\big)\right] \ldots\ldots\ldots (1)$$

This equation

1. Maximizes the probability of fake spectrum

Adversarial training

$$\min_D V(D) = -\mathbb{E}_{x\sim x}[\log D(x)] - \mathbb{E}_{y\sim y}\left[1 - \log D\big(G(y)\big)\right] \ldots\ldots\ldots\ldots (2)$$

This equation

1. Minimizes the probability of fake spectrum generated by the G network
2. Maximizes the probability of clean spectrum

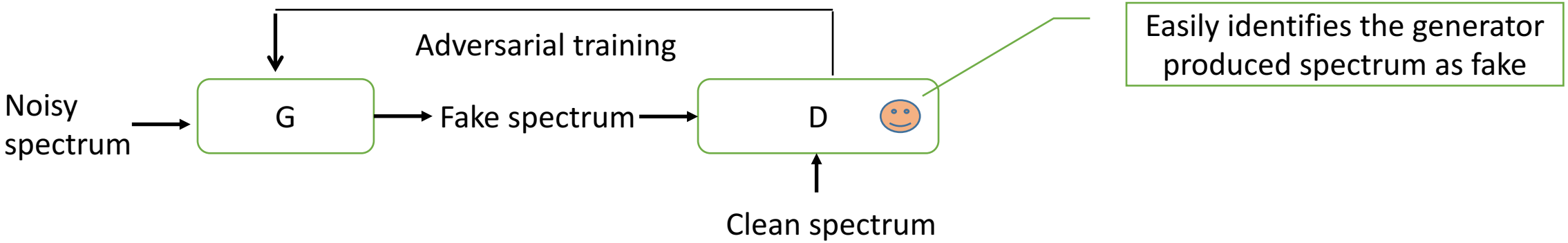
At the end the G network produces the enhanced spectrum.

Source:

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Bengio, Y, ”Generative adversarial nets”, in Advances in Neural Information Processing Systems *(NIPS)*, 2014, pp. 2672-2680.

Generative Adversarial Network (GAN)

Initially



After few epochs

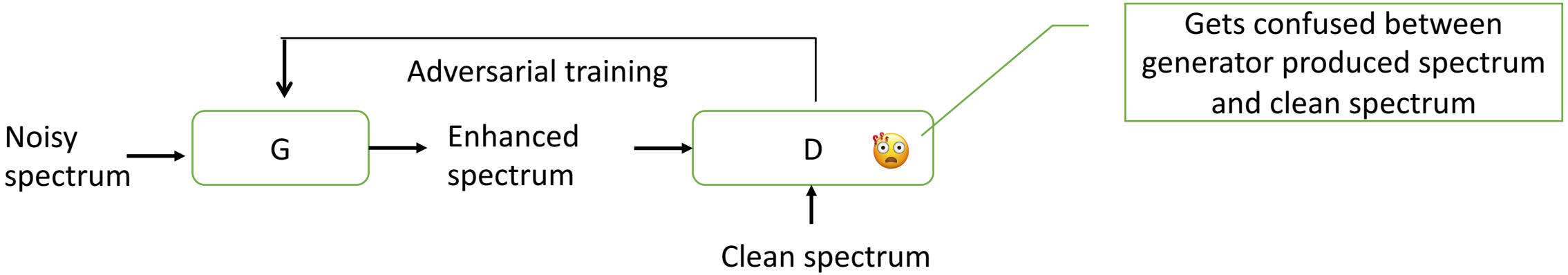


Fig 8. The training procedure employed in GAN.

T-F masking using Vanilla GAN (v-GAN)- The G network learns the mask implicitly and estimates the enhanced spectrum for the given noisy contextual spectrum

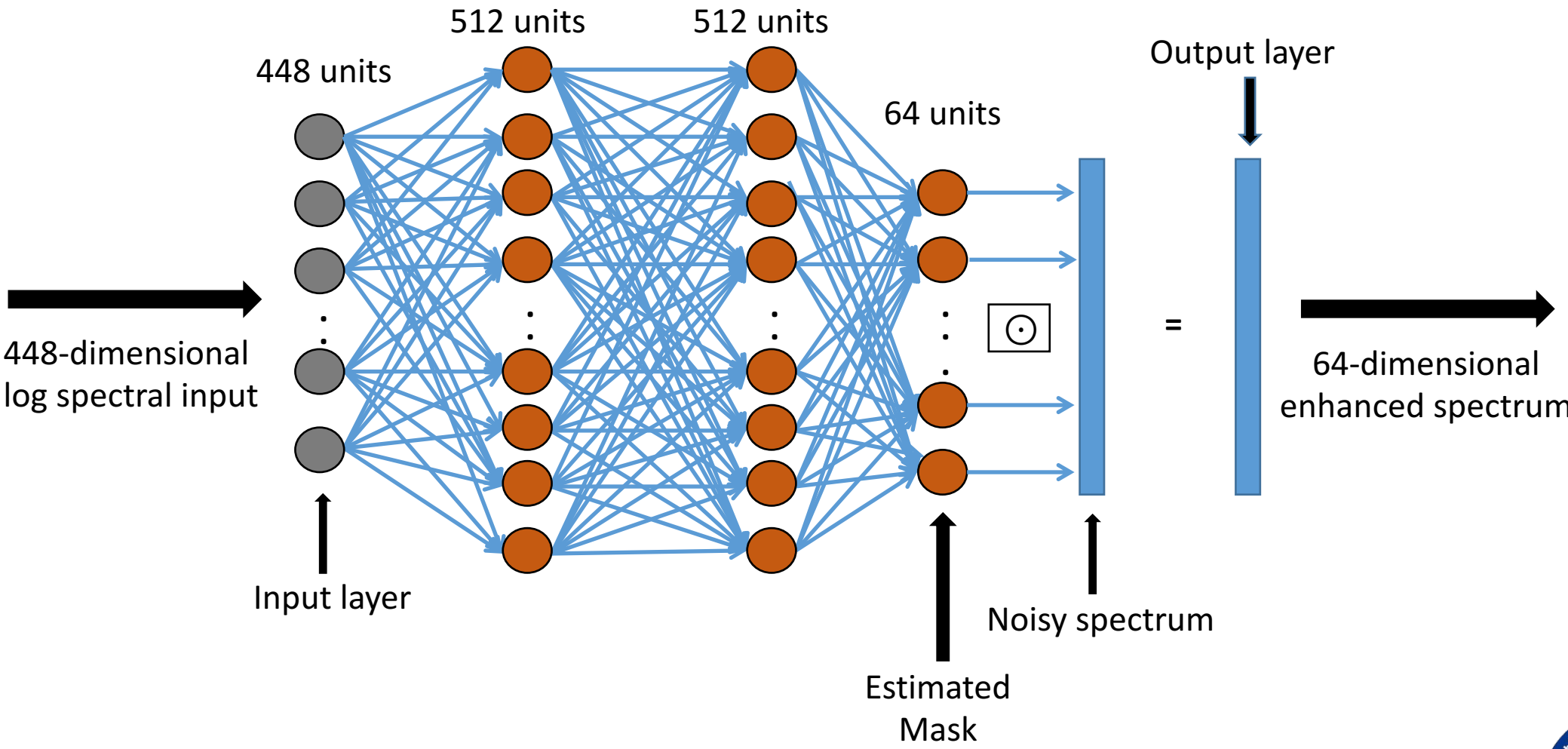


Fig 9. The generator network: inherently learnt mask and estimated enhanced spectrum.

T-F masking using Vanilla GAN (v-GAN) - Results

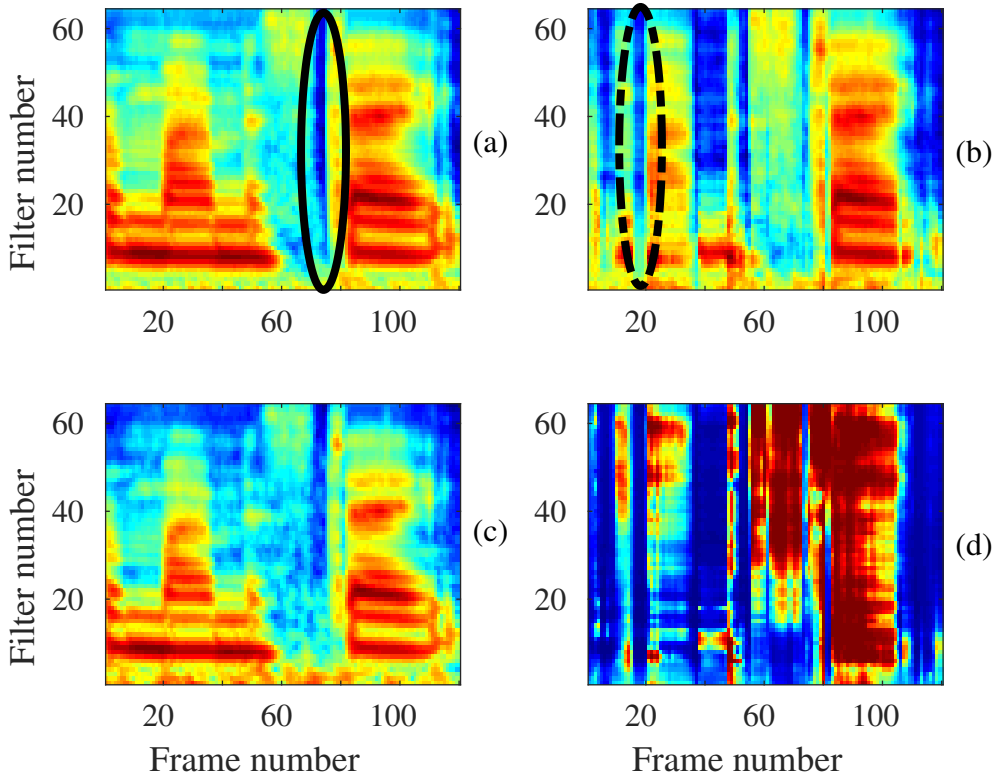


Fig 11. v-GAN fails to properly predict the mask: (a) Clean T-F representation: the solid-circle region shows the silence frame, (b) enhanced T-F representation: the dotted-circle shows the predicted frame where GAN fails, (c) noisy T-F representation and (d) predicted mask.

Source:

Meet H. Soni, Neil Shah, and Hemant A. Patil, ”Time-Frequency masking-based speech enhancement using Generative Adversarial Network”, to appear in the, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Alberta, Canada, 2018.

T-F masking using MMSE-GAN Objective functions for G and D network

Problem: The G network fools the D network by producing enhanced representation of some other frame.

Solution: Regularize the G network's objective function, by minimizing the Minimum Mean Square Error (MMSE) between the enhanced and the corresponding clean spectrum.

The D network's objective function remains the same

The modified G network's objective function is,

$$\min_G V(G) = -\mathbb{E}_{y\sim y}\left[\log D\big(G(y)\big)\right] + \frac{1}{2}\mathbb{E}_{x\sim x, y\sim y}\left[\log x - \log G(y)\right] \ldots\ldots\ldots (3)$$

$$\min_D V(D) = -\mathbb{E}_{x\sim x}[\log D(x)] - \mathbb{E}_{y\sim y}\left[1 - \log D\big(G(y)\big)\right] \ldots\ldots\ldots\ldots (4)$$

Source:

Meet H. Soni, Neil Shah, and Hemant A. Patil, "Time-Frequency masking-based speech enhancement using Generative Adversarial Network", to appear in the, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Alberta, Canada, 2018.

Network parameters for DNN, v-GAN, and MMSE-GAN

3 network are compared:

1. DNN
2. v-GAN
3. MMSE-GAN

Model	Input	3-Hidden layers	Output
DNN	448	512	64
G-network in GAN	448	512	64
D-network in GAN	64	512	1

Table 1. Selected parameters for the DNN, v-GAN, and MMSE-GAN architecture

- 64-channel Gammatone filterbank with 20 *ms* Hamming window length and 10 *ms* window shift, and 7 frame context
- ADAM optimizer with learning rate 0.001 and batch size of 1000

Source:

Meet H. Soni, Neil Shah, and Hemant A. Patil, ”Time-Frequency masking-based speech enhancement using Generative Adversarial Network”, to appear in the, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Alberta, Canada, 2018.

Database used

- The database released by Valentini et. Al. is used for evaluating the algorithm
- The training and testing set have mismatched conditions
- The noisy training set is prepared with a total of 40 different noisy conditions with 10 types of noise and 4 signal-to-noise ratio (SNR) each (15, 10, 5, and 0 dB)
- The noisy test set is prepared with a total of 20 different noisy conditions with 5 types of noise and 4 signal-to-noise ratio (SNR) each (17.5, 12.5, 7.5, and 2.5 dB)
- The database comprises of 11572 training utterances and 824 testing utterances

Source:

B. Valentini, Cassia, et al. ”Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech,” in 9th ISCA Speech Synthesis Workshop, Sep. 13-15, Sunnyvale, CA, USA, 2016. http://datashare.is.ed.ac.uk/handle/10283/1942/, [online; Last Accessed 25-July-2017].

Select the best model that gives least MSE on the validation set

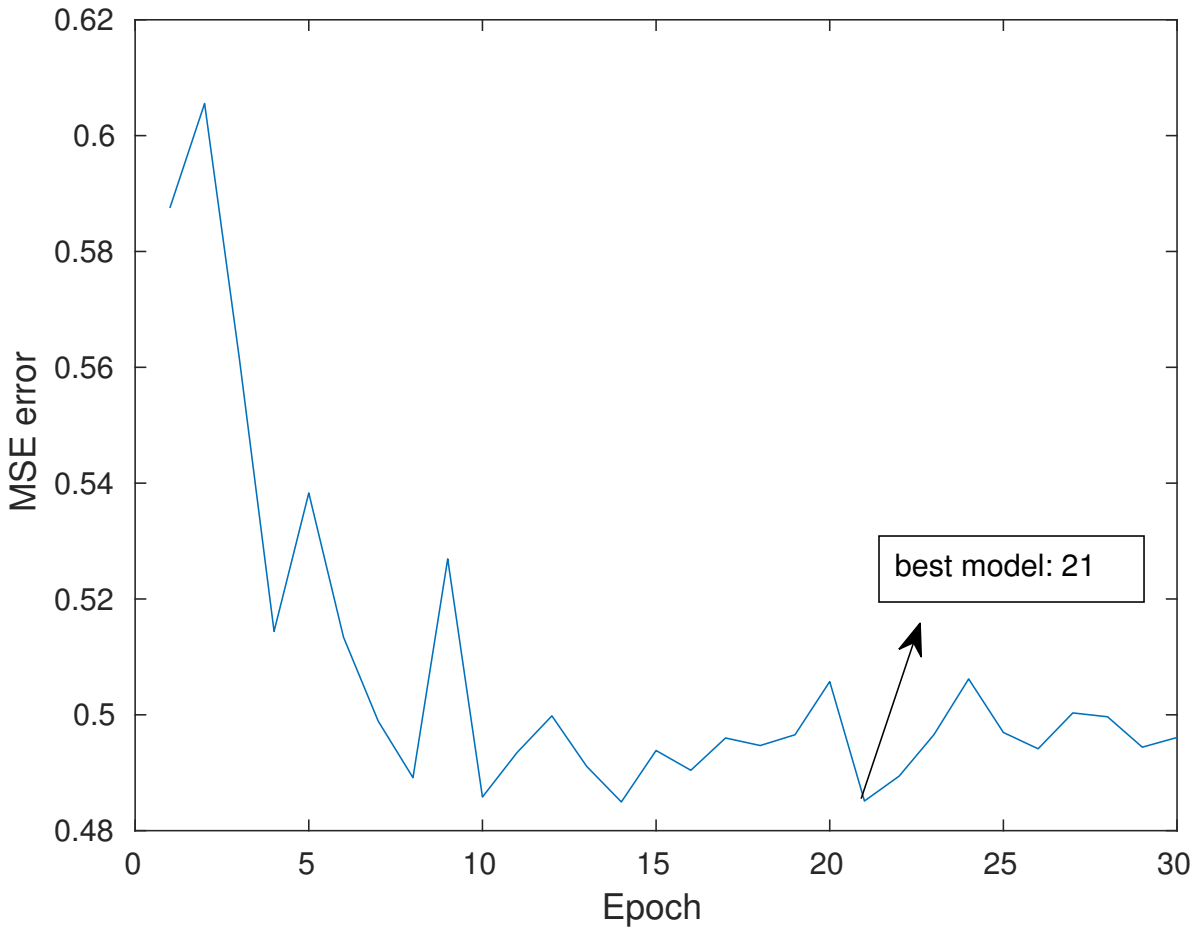
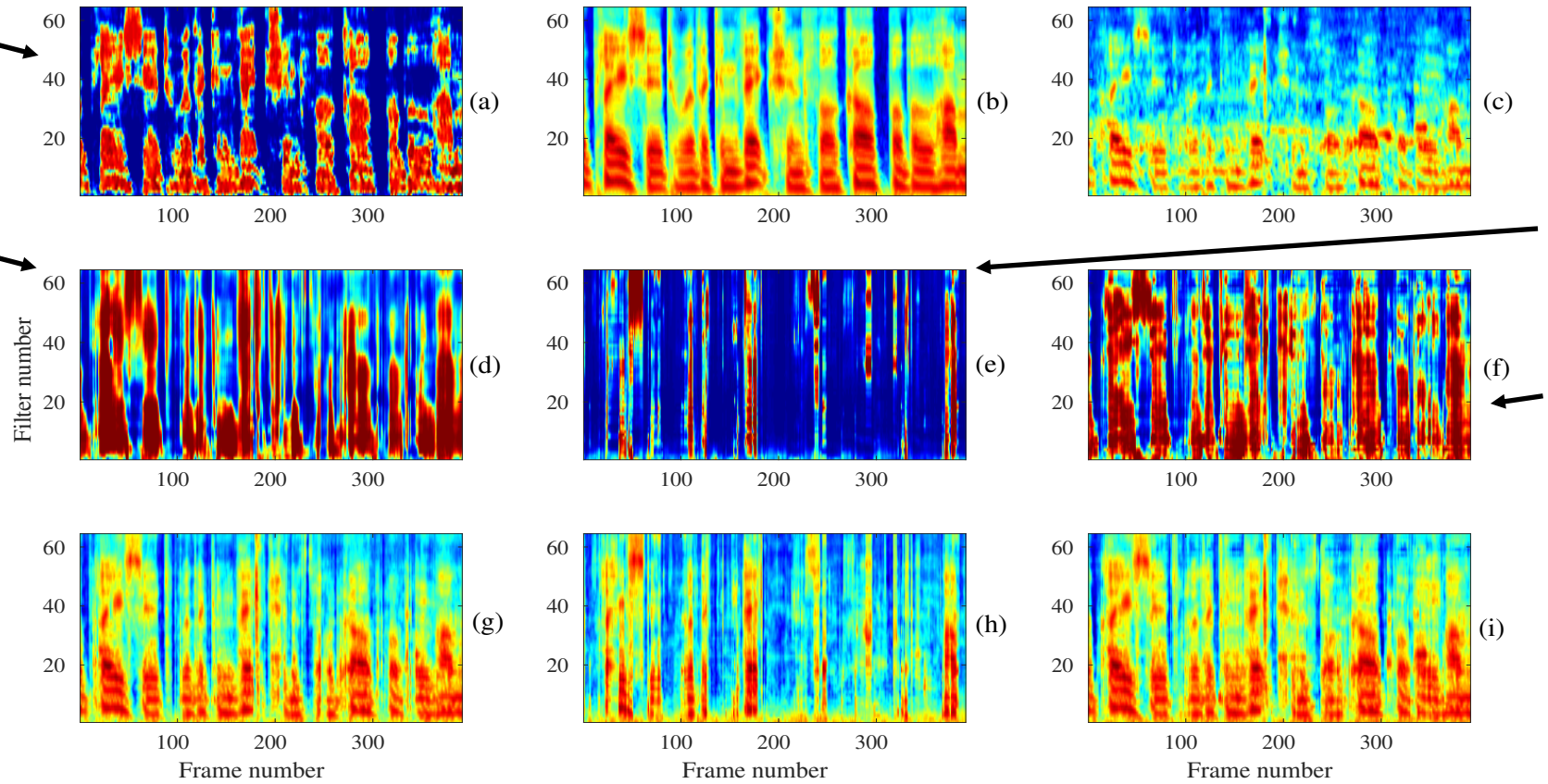


Fig 12. MSE loss *vs.* training epochs

Results of T-F masking using DNN, v-GAN, and MMSE-GAN architecture

Oracle mask

DNN: Better than v-GAN, not better than MMSE-GAN



v-GAN: Very poor mask prediction by v-GAN

MMSE-GAN: Intelligible predicted mask

Fig 13. (a) Oracle mask, Gammatone spectrum of (b) clean speech, (c) noisy speech. Predicted mask using (d) RMSE-DNN, (e) GAN, (f) RMSE-GAN. Gammatone spectrum of reconstructed speech using (g) DNN, (h) GAN, (i) MMSE-GAN.

Source:

Meet H. Soni, Neil Shah, and Hemant A. Patil, ”Time-Frequency masking-based speech enhancement using Generative Adversarial Network”, to appear in the, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Alberta, Canada, 2018.

Performance measures- Objective scores

1. **Short-Time Objective Intelligibility (STOI)** (0-1) and **Perceptual Evaluation of Speech Quality (PESQ)** (-0.5 to 4.5) denotes the correlation between the clean and enhanced speech.
2. **Composite measure for signal distortion (CSIG)**
3. **Composite measure for background interferences (CBAK)**
4. **Composite measure for overall speech quality (COVL)**

Source:

- P.862.2: Wideband extension to Recommendation P.862 for the assesement of wideband telephone networks and speech codecs, ITU-T Std. P.862.2, 2007.
- Y. Hu and P.C.Loizou, ”Evaluation of objective quality measures for speech enhancement,” IEEE Trans. on Audio, Speech, and Language Processing (TASLP), vol. 16, no. 1, pp. 229-238, Jan 2008.

Results of T-F masking using DNN, v-GAN, and MMSE-GAN architecture

Metric	Noisy	DNN	v-GAN	MMSE-SEGAN	SEGAN	Wiener
CSIG	3.35	3.73	2.48	**3.80**	3.48	3.23
CBAK	2.44	3.09	2.64	**3.12**	2.94	2.68
CMOS	2.63	3.09	1.91	**3.14**	2.8	2.67
PESQ	1.97	2.49	1.41	**2.53**	2.16	2.22
STOI	0.91	**0.93**	0.79	**0.93**	**0.93**	-

Table 2. Performance comparisons between the noisy signal, DNN, v-GAN, MMSE-GAN , SEGAN and the Wiener filter-based enhancement.

Source:

- Meet H. Soni, Neil Shah, and Hemant A. Patil, ”Time-Frequency masking-based speech enhancement using Generative Adversarial Network”, accepted in the, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Alberta, Canada, 2018.
- Pascual, Santiago, Antonio Bonafonte, and Joan Serra. "SEGAN: Speech enhancement generative adversarial network”, in Proc. of the Interspeech, Stockholm, Sweden, 2017, pp. 3642-3646.

Results of T-F masking using DNN, v-GAN, and MMSE-GAN architecture

1. v-GAN follows the same objective function as that of the traditional GAN.
2. MMSE-GAN simply modifies v-GAN objective function by adding a MMSE regularizer.
3. The MMSE-GAN architecture leads to an improved performance over DNN, which is the state-of-the-art SE technique.
4. Comparison with SEGAN (INTERSPEECH 2017) suggests that **T-F masking-based approach is better for SE task.**

Source:

Meet H. Soni, Neil Shah, and Hemant A. Patil, ”Time-Frequency masking-based speech enhancement using Generative Adversarial Network”, to appear in the, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Alberta, Canada, 2018.

Speech Research Lab

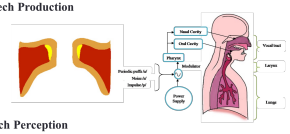
Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, INDIA.

Speech and language are the most powerful means of communication between human beings. Speech carries different levels of information such as linguistic message content, speaker's identity, his/her health, emotions, attitude, gender, age, acoustic environment in which it is recorded. The aim of Speech Lab at DA-IICT group is to use this information hidden in speech and create an interface between human speech and computer for certain important tasks such as speech and speaker recognition, analysis of voice biometric attacks, speech synthesis, audio search applications, voice conversion, etc. The group is currently focusing on algorithms and systems of speech processing. The aim of group is to provide the researchers the freedom to work together in pursuing the ideas with passion and enthusiasm. The lab is mentored by Prof. Hemant A. Patil.

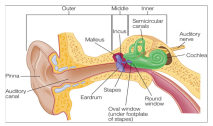
Current Members

Speech Production and Perception

1. Speech Production



2. Speech Perception

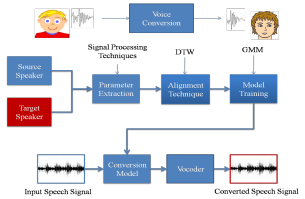


Summary of Activities

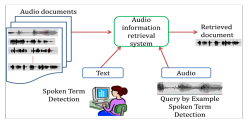
- Lab produced **03 Ph.D.** thesis and **37 M.Tech.** thesis and currently 05 doctoral scholars and 05 masters students are working
- **200** research publications
- Lab is executed 03 sponsored projects (and 02 ongoing) of worth **2.25 crores.**
- Lab has state-of-the-art equipment, speech corpora, servers.
- Lab alumni has secured position in industry (TCS Innovation Labs, IBM, Infosys, Deloitte, etc.), academia (UT Dallas, NUS Singapore, EURECOM, France, Baidu Speech Research, Microsoft Research, Bangalore, IISc Bangalore, etc.)
- Lab members presented papers in top conferences (ICASSP, INTERSPEECH, EUSIPCO, etc.)
- Lab members have received grants from IEEE SPS, Microsoft, IBM, ISCA, etc. to present research papers

Research Activities

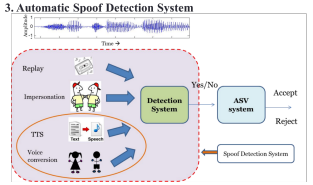
1. Voice Conversion System



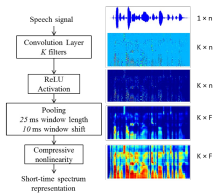
2. Spoken Term Detection



3. Automatic Spoof Detection System

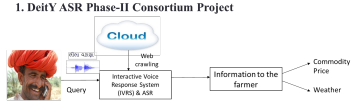


4. Representation Learning for Speech Processing

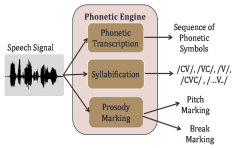


Research Projects

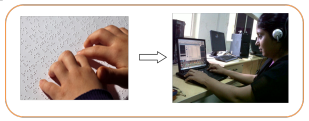
1. DeitY ASR Phase-II Consortium Project



2. DeitY Prosody Phase- II: Phonetic Engine Project



3. DeitY TTS Phase-II Project: Application of for visually challenged people



4. Acoustic Analysis of Musical pillar analysis at Vitthala Temple, Hampi (DST-IDH Hampi Project)

5. Infant Cry Analysis (DST FastTrack Award to Dr. Hemant A. Patil)

Achievements

1. International Recognition by ISCA

2. Co-edited Book

3. DA-IICT TTS Team Topped in ASV Spoof Challenge 2015, INTERSPEECH 2015, Dresden, Germany, Sept. 2015.

4. Best Paper Award

5. E-content Development for Digital India Program

6. Prof. Patil received NVIDIA Hardware Grant Titan X GPU for Speech Research

Speech Research Lab, DA-IICT, Gandhinagar, India.

Thank you