

Deep Learning for Speech Enhancement

Laurent Girin

Univ. Grenoble Alpes / Grenoble-INP / GIPSA-lab / INRIA

September 5, 2018

Speech enhancement: The problem

- $x(\tau) = s(\tau) + n(\tau) \rightarrow \text{process} \rightarrow \hat{s}(\tau)$
- Goal: improve the quality and intelligibility of speech for telecommunication systems, improve ASR scores
- An old topic of signal processing, renewal coming together with the arrival of home assistants / smart loudspeakers (Amazon Echo, Google Home, Apple Homepod...)
- Traditionally we distinguish the single-channel case and the multi-channel case
- Solution(s): 50 years of signal processing literature...



Speech enhancement: The problem

- $x(\tau) = s(\tau) + n(\tau) \rightarrow \text{process} \rightarrow \hat{s}(\tau)$
- Goal: improve the quality and intelligibility of speech for telecommunication systems, improve ASR scores
- An old topic of signal processing, renewal coming together with the arrival of home assistants / smart loudspeakers (Amazon Echo, Google Home, Apple Homepod...)
- Traditionally we distinguish the single-channel case and the multi-channel case
- Solution(s): 50 years of signal processing literature...
- ...and #5 years of amazing new developments with a deep learning approach that changed it all!



Speech enhancement: The problem

- $x(\tau) = s(\tau) + n(\tau) \rightarrow \text{process} \rightarrow \hat{s}(\tau)$
- Goal: improve the quality and intelligibility of speech for telecommunication systems, improve ASR scores
- An old topic of signal processing, renewal coming together with the arrival of home assistants / smart loudspeakers (Amazon Echo, Google Home, Apple Homepod...)
- Traditionally we distinguish the single-channel case and the multi-channel case
- Solution(s): 50 years of signal processing literature...
- ...and #5 years of amazing new developments with a deep learning approach that changed it all!
- Please, have a look at Emmanuel Vincent's WASPAA 2015 Keynote Talk slides "Is audio signal processing still useful in the era of machine learning?"



Old school signal processing approaches (1970-2010)²

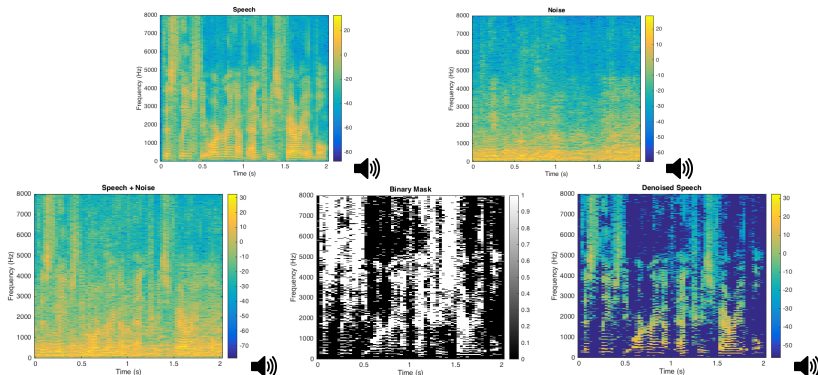
- Let us start with the single-channel case
- Typically an estimation problem in the STFT domain
- Historical method 1: Estimate the short-term power spectral density of noise $\gamma_{n,ft} = E[|n_{ft}|^2]$ during speech silence, then apply spectral subtraction $\hat{\gamma}_{s,ft} = \max(|x_{ft}|^2 - \hat{\gamma}_{n,ft}, 0)$ and Wiener filtering $\hat{s}_{ft} = \frac{\hat{\gamma}_{s,ft}}{\hat{\gamma}_{s,ft} + \hat{\gamma}_{n,ft}} x_{ft}$
- Historical method 2: Bayesian estimation of short-term spectral amplitude $|s_{ft}|$ ¹
- Speech and noise must have different spectral characteristics. Usually, noise is assumed more stationary than speech.
- Single-channel *multispeaker* separation is extremely difficult.

¹Y. Ephraim and D. Malah. "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32.6 (1984).

²P. Loizou. *Speech enhancement: theory and practice*. CRC press, 2007. 

The CASA approach (mid 90's-2010's)³

- Estimate a binary mask $M_{ft} = 1$ if TF-point $\{ft\}$ is dominated by speech, and $M_{ft} = 0$ if it is dominated by noise
- Then, $\hat{s}_{ft} = M_{ft} \cdot x_{ft}$



- Note: Extensions to soft masks ($0 \leq M_{ft} \leq 1$), related to Wiener filters

³DL Wang and G.J. Brown. *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE press, 2006.

The CASA approach (mid 90's-2010's)

- In practice the mask is estimated from the noisy speech spectrogram. You must identify and group TF points with speech characteristics (e.g. harmonicity, common onsets/offsets/modulations.) This is done with a mixture of techniques involving statistical signal processing, speech analysis, human audition modeling, computer vision and pattern analysis, etc.
- This is a very difficult task in single-channel configuration, especially when noise is strong.
- Not suited for multispeaker separation.
- Relatively “poor” results (compared to ideal binary masking); Limited quality of enhanced signals (musical noise artefacts); Not very well suited for ASR.

Deep Learning approach – Revolution I (2013-2015)⁴

- General principle: Speech enhancement is turned into a supervised/data-driven regression problem using deep artificial neural networks
- Basically, two “lines of research”:
 - Regression from noisy speech spectrogram to clean speech spectrogram + use of the phase of noisy signal
 - Regression from noisy speech spectrogram to mask + application of the estimated mask to noisy signal
- Use of Feedforward-DNNs, CNNs, LSTMs, etc.
- Requires a huge amount of parallel clean and noisy training data (is that really a problem these days?)
- New standards of performances – Typically 8–12dB output SNR = +3–10dB compared to CASA / NMF, etc. (Note: Mask estimation works a bit better than direct estimation of clean spectra).

⁴DL Wang and J. Chen. “Supervised speech separation based on deep learning: an overview”. In: *arXiv preprint arXiv:1708.07524* (2017).

Deep Learning approach – Revolution I

- This acted as an electroshock in the audio processing community: DL can solve a 100% signal processing problem!

Deep Learning approach – Revolution I

- This acted as an electroshock in the audio processing community: DL can solve a 100% signal processing problem!

Decades of development of signal processing / CASA machinery shortly replaced with a data-driven blackbox!



Deep Learning approach – Revolution I

- This acted as an electroshock in the audio processing community: DL can solve a 100% signal processing problem!

Decades of development of signal

processing / CASA machinery shortly

replaced with a data-driven blackbox!



- Deep approach but shallow (and boring) science. TONS of papers with DNN regression, discussing the effects of different DNN models, i/o data representations, training criteria, datasets, etc., often leading to more or less the same results. This is really the dark side of DL (research), is it not??!!

Deep Learning approach – Revolution I

- This acted as an electroshock in the audio processing community: DL can solve a 100% signal processing problem!

Decades of development of signal

processing / CASA machinery shortly

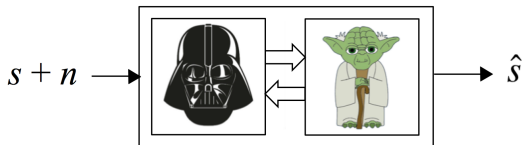
replaced with a data-driven blackbox!



- Deep approach but shallow (and boring) science. TONS of papers with DNN regression, discussing the effects of different DNN models, i/o data representations, training criteria, datasets, etc., often leading to more or less the same results. This is really the dark side of DL (research), is it not??!!
- Explanations for this success exist! e.g. DNNs are powerful models that can account for complex (non-linear) dependencies of data across TF points and are highly scalable with data size, whereas most traditional SP techniques assume (conditional) independence of data across TF points and are poorly scalable with data size.

After the revolution

- General goal: Put the best of both worlds together!
→ Combine DL models with conventional statistical signal processing, generative models, Bayesian estimation, CASA...



- Example: Variational Autoencoder (VAE) speech model + Gaussian-NMF noise model.^{5,6} VAE is trained offline, but estimation of NMF parameters at test time → can adapt to any kind of noise. Enhancement from 0dB iSNR \llcorner to ≈ 14 dB oSNR \llcorner in ⁶ (semi-superv. NMF baseline ≈ 11 dB \llcorner).

⁵Y. Bando et al. "Statistical Speech Enhancement Based on Probabilistic Integration of Variational Autoencoder and Non-Negative Matrix Factorization". In: *arXiv preprint arXiv:1710.11439* (2017).

⁶S. Leglaive, L. Girin, and R. Horaud. "A variance modeling framework based on variational autoencoders for speech enhancement". In: *IEEE Int. Workshop on Machine Learning for Signal Processing (MSLP)*. 2018.

Deep learning approach – Revolution II: Single-channel **multispeaker separation**

- In single-channel configuration, **separation of two speakers** (and more) is much more difficult than speech enhancement in noise because the interfering signal is here of the same nature as the target signal (actually, it is another target signal). So how can a network (or CASA model) learn two different models of the same thing?

Single-channel multispeaker separation

One breakthrough technique: Deep Clustering^{7,8}

- Supervised technique using training pairs of speech signals and their summation
- Huge BLSTM network trained to project each point of the power spectrogram of the mixture signal (for a complete utterance) into a space of *embedding vectors* that are easier to separate
- At training time, embeddings vectors corresponding to TF points dominated by the same speaker are forced to be aligned whereas embeddings vectors corresponding to TF points dominated by a different speaker are forced to be orthogonal. At testing time, clustering of the embedding vectors leads to efficient separation (which is done with mask).
- Unprecedented separation performance for a two-speaker single-channel mixture – Typically up to 15dB output SNR, close to ideal TF masking.
- See the demo presented at SANE 2017, available on Youtube.

⁷J. R. Hershey et al. "Deep clustering: Discriminative embeddings for segmentation and separation". In: *IEEE ICASSP*. 2016.

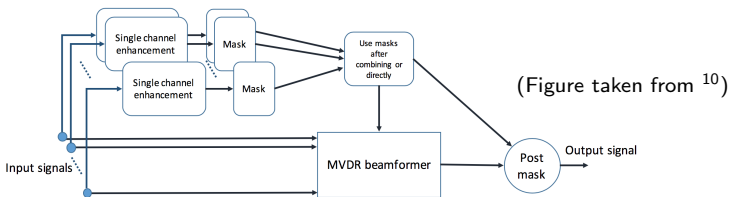
⁸Y. Isik et al. "Single-channel multi-speaker separation using deep clustering". In: *arXiv:1607.02173* (2016).

And now for the multichannel case

- Multichannel recordings provide i) more signal samples (good to remove noise), and ii) spatial information related to relative position of the sound source(s) w.r.t. the mic-array
- Conventional signal processing: *Beamforming*⁹ = Spatial information is used to build enhancing spatial filters, e.g. MVDR. Possibly combine with estimated spectral info on the source – **which is still a difficult task** – e.g. MWF.
- CASA: Spatial information is used to cluster the TF points into different sources and then estimate binary masks again
- Improvement of performance compared to single-channel case. Good separation for a reasonable amount of sources, spatial separation, and reverberation. **Speech denoised with binary mask remains poorly suited for ASR.**

⁹B. D. Van Veen and K. M. Buckley. "Beamforming: A versatile approach to spatial filtering". In: *IEEE ASSP Magazine* 5.2 (1988), pp. 4–24.

Deep learning based multi-channel speech enhancement: First approach



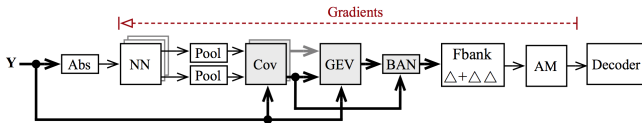
- “Basic” combination of DL-based single-channel speech enhanc. (spectral info) and beamforming (spatial info):^{10,11,12} DNNs provide masks, used to select s/n points, used to estimate s/n covariance matrices, used to build BF filters.
- Better ASR scores than with “direct” masking and basic BF.

¹⁰H. Erdogan et al. “Improved MVDR Beamforming Using Single-Channel Mask Prediction Networks”. In: *INTERSPEECH*. 2016.

¹¹J. Heymann, L. Drude, and R. Haeb-Umbach. “Neural network based spectral mask estimation for acoustic beamforming”. In: *IEEE ICASSP*. 2016.

¹²T. Higuchi et al. “Deep Clustering-Based Beamforming for Separation with Unknown Number of Sources”. In: *Interspeech*. 2017.

Deep learning based multi-channel speech enhancement: First approach, end-to-end extension






- Joint optimization of the mask estimator, BF, and ASR AM.¹³
- Note: For ASR application, you do not need to regenerate the output enhanced speech signal

¹³ J. Heymann et al. "BEAMNET: End-to-end training of a beamformer-supported multi-channel ASR system".
In: *IEEE ICASSP*. 2017.

Deep learning based multi-channel speech separation: Second approach¹⁵

- Combination of DNN-based single-channel spectral enhancement + source separation technique based on Spatial Covariance Matrix model and Wiener filtering.¹⁴

These two processes are iterated in an EM-like algorithm, typical of many generative model approaches.

- Evaluation: CHIME-3 data. SDR: 13.25dB and WER = 10.1% on simu test data, vs. 7.72dB and 13.4% with NMF-EM baseline.
- Noisy signal  NMF  DNN-SCM 

¹⁴N. Q. K. Duong, E. Vincent, and R. Gribonval. "Under-determined reverberant audio source separation using a full-rank spatial covariance model". In: *IEEE Trans. Audio, Speech, Lang. Process.* 18.7 (2010), pp. 1830–1840.

¹⁵A. A. Nugraha, A. Liutkus, and E. Vincent. "Multichannel audio source separation with deep neural networks". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.9 (2016), pp. 1652–1664.

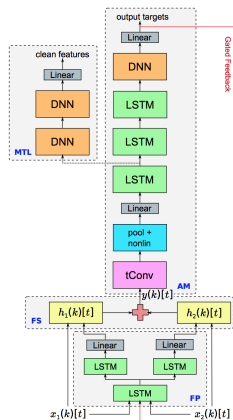
Deep learning based multi-channel speech enhancement:

Third approach

- Direct estim. of BF filters with DNNs + end-to-end training
- Time-Domain implementation (!) (Google):¹⁶

LSTM network takes raw multichannel signal input and outputs multichannel **5-ms** FIR filters. TD filtering & sum produces single-channel enhanced signal, passed to a “raw waveform AM” (Convolutional LSTM which actually include some kind of filterbank). End-to-end training of filter predictor and AM. Training with 2,000h of data, room sim. speech+ambient noise (SNR in 0-20dB), many source/sensor/room configurations, but only 2-channel set-up (!). WER \approx 20%.

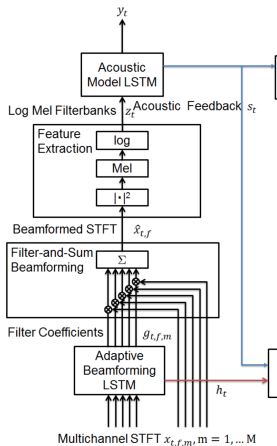
Improvement over baseline is marginal.



Deep learning based multi-channel speech enhancement: Third approach

- Direct estim. of BF filters with DNNs + end-to-end training
- TF-domain implementation (MERL):¹⁷

LSTM network takes multichannel STFT input and outputs adaptive multichannel STFT filters (at frame level) (concatenation of real and imag. parts, and freq. and channels). Filter & Sum BF. Logmel spectrum. Deep LSTM-HMM AM. Feedback of AM decoding at previous frame to input. End-to-end training of BF network and AM (after sequential training of components). Evaluation on CHIME-4 data. WER \approx 20 – 25%. Improvement over AM alone, **but NOT over Basic BF+AM on real data!**



Conclusion

- DL is totally amazing!!!

Conclusion

- DL is totally amazing because it allows you to make a presentation at a workshop organised by maths institutions without presenting a battery of stylish equations!

Conclusion

- DL is totally amazing because it allows you to make a presentation at a workshop organised/supported by maths institutions without presenting a battery of stylish equations!
- About the DL-based speech enhancement literature:
Good papers 😊 and bad papers 😞
- Breakthrough performances with initial “simple” application of DNNs to the speech enhancement problem. So far, and in a very general manner, results obtained with more and more sophisticated combinations of DL and SP techniques are poorly improved compared to basic combinations.
- We can do better. DNNs have great modeling power. Domain knowledge and generative models have proven to be useful. So, let us pursue the quest for combining the best of both worlds into smart hybrid DNN/SP architectures! There is still plenty of room for new developments.