# Speech Separation in Supervised Setting

Sravan Patibandla Jaideep Patel Mrinmoy Maity

## 1 MOTIVATION

Source separation of audio signals is relevant in numerous real world application. We humans are trained to separate speech from noise in order to understand the information it carries. However, training machines to acquire the same capability is a challenging task. Every single source of signal has its own characteristics which can be combined with a variety of noises to give huge variation on which the model has to operate on.

Designing a generalized model that can handle a variety of noisy signal and filter out relevant sounds can be difficult. But if it can be achieved, the artificially trained model can be used in a number of tasks. For example, separating background claps from a music recorded in a concert lets us understand the lyrics properly if we want to hear it later. Or it may allow us to compare the performance of singer compare to other concerts the singer has performed or even its corresponding studio version. Or if one listens to a song and want to search the same song, its important to filter out background noise in order to be able to increase the accuracy of search result. Same concept can be applied to speech where recorded conversation on a busy street becomes automatically filtered when using the model. With the ubiquitous use of phones, one may also want to filter out background noise to send over clean signals over the network while talking in midst of crowded park. This would require the model to be efficient so that filtering can happen instantaneously.

The problem of source separation comes in all forms: from multi-speaker to single-speaker multiple recordings to effect of room reverberations and spatial source configuration on speech. Monaural speech recognition is a class of problems which falls under this broad category and of special interest to researchers because of few reasons. Monaural speech separation means speech separation from single microphone recordings and much less susceptible to

room reverberation. Because of its simplicity yet wide applicability, this class of problems has been widely studied among researchers.

## 2 Related Works

Human auditory system performs pretty well in monaural speech separation. In [5], based on Auditory Scene Analysis, our auditory system separates an auditory signal into multiple streams, each corresponding to one sound source. The process works in two independent stages: first the signal is decomposed into segments and then based on periodicity, the segments coming from same sources are grouped.

Speech separation techniques have been quite well studied and most of the methods falls in two major categories: signal processing based and model based. Signal processing based models operate under the assumptions of speech and noise distributions. This approach has limited performance in low signal-to-noise ratio. Statistical model-based methods[13] infer speech spectral coefficients given noisy observations under prior distribution assumptions for speech and noise. Non-negative matrix factorization method[12] models noisy observations as weighted sums of non-negative source bases. But they do not generalize well to unseen noisy conditions and are mostly effective for structured interference. Model based methods in [6] overcomes this limitation and performs reasonably well in low SNR conditions.

In this work, monaural speech separation has been formulated using a data driven approach: a supervised learning problem. A model will be provided representation of noisy signal and trained on some representation of a clean signal. Weiner filter is considered to be an optimal filter to recover noisy speech. Our work is largely based on methods and experiments proposed in [2]. Section 3 will discuss about the basic framework adopted for monaural speech separation. In section 4, we will discuss the experimental setup and results obtained. Section 5 will conclude our findings and discuss about future scope of the project.

## 3 Model Framework and Targets

We can visualize our model as a deterministic non-linear mapping function $f : N \rightarrow S$ where $f$ maps representation of noisy signal $N$ to representation of clean speech $S$. Speech signal are continuous valued variable which is represented as a function of amplitude and phase. To simplify mapping function $f$, we consider only amplitude and ignore the impact of phase on recovery. Representation of signals can be obtained using features like mel-frequency central coefficients(MFCC), amplitude modulation spectra(AMS) or simply spectrogram representation of continuous wave. In our experiments, we have used amplitudes of spectrum assuming

constructed model will learn from data itself without explicit feature engineering.

There are two ways we can approach this problem: Either model architectures can be changed keeping the same target variable to learn the recovery process or same model can be used for different targets. Here we took the latter approach where by keeping the model architecture fixed, we present an empirical results to provide a comparative analysis of different targets. Deep Neural Networks has been used in various applications like computer vision[7,8], natural language processing[9,10]. Its popularity is mostly based on its application agnostic nature and data driven approach.

We employ a comparatively shallower version of feedforward multilayer perceptron for our speech denoising problem for faster inference. All trained networks consists of three hidden layers with 1024 units each and ReLU is used for activation in all units in all the layers. To prevent overfitting, we used regularization in form of dropout. As proposed in [3], dropout rate 0.1 applied to input layer and 0.5 to all hidden layers for optimum performance. In our experiments, a small variation of dropouts did not have a significant performance degradation. The architecture is learnt using stochastic gradient descent with momentum to fasten the convergence of loss function, mean squared error. The dimensionality of the target function is same as input.

The goal of auditory perception is highly subjective. Like computer vision where the purpose of segmentation and object detection is to make sense of surrounding environment, auditory sense also highly dependent on listeners and their ability to hear the distinctness of speech or melody of instrumental music. This in term allows us to choose different kinds of targets to train the model. And that makes finalizing on the type of training target difficult. In following subsections , we will discuss some of widely used targets to train DNN in a supervised setting.

### 3.1 Ideal Binary Masks

A widely used representation of signal is its corresponding time-frequency(T-F) mapping where time representing time slices in original signals with or without overlaps and frequency signifies the auditory filter bank being able to perceive by humans. Ideal binary mask represents a binary matrix where 1 signifies speech signal is stronger than interference signal and 0 means the noise signal is supersedes speech for each time-frequency cell. The use of ideal binary masks is well argued in computational auditory scene analysis[4]. While training our model, we use element wise sigmoid function to limit the output range in [0,1] as the mask itself is binary. Leaving out this stage increases the error rate and makes the model parameters difficult to converge. Rounding of models output to nearest integer 0,1 gives us the estimated ideal binary mask.

$$IBM(t,f) = \begin{cases} 1 & \text{if } S(t,f) > N(t,f) \\ 0 & otherwise \end{cases}$$

Table 4.1: Test Result (STOI) for -5 db SNR for Male Speakers. Trained on both genders

| | birds | jungle | motorcycles | ocean | keyboard | machinegun |
|---|---|---|---|---|---|---|
| IBM | 0.72 | 0.66 | 0.62 | 0.58 | 0.79 | 0.74 |
| IRM | 0.38 | 0.38 | 0.56 | 0.44 | 0.60 | 0.66 |

Table 4.2: Test Result (STOI) for 0 db SNR for Male Speakers. Trained on both genders

| | birds | jungle | motorcycles | ocean | keyboard | machinegun |
|---|---|---|---|---|---|---|
| IBM | 0.81 | 0.76 | 0.74 | 0.71 | 0.84 | 0.79 |
| IRM | 0.53 | 0.53 | 0.68 | 0.58 | 0.65 | 0.70 |

### 3.2 Ideal Ratio Masks

Sometimes we want to avoid the binarize or rounding up and instead want to inspect the extent to which interference signal mixes with clean signal. The difference of IRM with IBM is that we are target variables falls in the range of [0,1] instead of 0,1. IRM is defined as,

$$IRM(t,f) = \left(\frac{S^2(t,f)}{S^2(t,f)+N^2(t,f)}\right)^{\beta} \tag{3.1}$$

It also allows to introduce a tunable parameter $\beta$ to scale the mask. As referred in [2], optimal parameter is found to be $\beta$=0.5 which also makes IRM to be similar to square-root of Weiner filter.

## 4 Evaluation & Experimental Results

The models has been trained on 8core CPU with 52 GB internal memory. To expedite training time, we took help of 2 x NVIDIA Tesla K80 GPUs, 12 GB memory each on Google Compute Engine. The experiment has been done on TIMIT dataset . The dataset contains 8 different dialects with each dialect contains male and female speeches. We have noise dataset mimicking real world noises like birds, keyboard, ocean etc. Due to its scale and larger processing time, we experimented on a sample of the original dataset. First, 10 speeches of males and female are selected independently and mixed with 5 different noises at 3 scales(-5db, 0db and 5db). The noise are extracted choosing a random starting offset. To verify our setup, each gender voice, noise and a scale are saved in files. The file has a dictionary with keys X,S,N representing noisy, clean speech and noise respectively. Each Key has a list of spectrograms

Table 4.3: Test Result (STOI) for 5 db SNR for Male Speakers. Trained on both genders

| | birds | jungle | motorcycles | ocean | keyboard | machinegun |
|---|---|---|---|---|---|---|
| IBM | 0.86 | 0.84 | 0.82 | 0.81 | 0.88 | 0.86 |
| IRM | 0.66 | 0.66 | 0.75 | 0.69 | 0.70 | 0.74 |

Table 4.4: Test Result (STOI) for -5 db SNR for female Speakers. Trained on both genders

| | birds | jungle | motorcycle | ocean | keyboard | machinegun |
|---|---|---|---|---|---|---|
| IBM | 0.66 | 0.59 | 0.59 | 0.47 | 0.76 | 0.71 |
| IRM | 0.33 | 0.34 | 0.58 | 0.41 | 0.6 | 0.67 |

Table 4.5: Test Result (STOI) for 0 db SNR for female Speakers. Trained on both genders

| | birds | jungle | motorcycle | ocean | keyboard | machinegun |
|---|---|---|---|---|---|---|
| IBM | 0.76 | 0.7 | 0.7 | 0.65 | 0.81 | 0.77 |
| IRM | 0.47 | 0.51 | 0.69 | 0.55 | 0.65 | 0.71 |

Table 4.6: Test Result (STOI) for 5 db SNR for female Speakers. Trained on both genders

| | birds | jungle | motorcycle | ocean | keyboard | machinegun |
|---|---|---|---|---|---|---|
| IBM | 0.82 | 0.78 | 0.77 | 0.76 | 0.85 | 0.81 |
| IRM | 0.62 | 0.64 | 0.74 | 0.66 | 0.7 | 0.74 |

Table 4.7: Test Result (STOI) for -5 db SNR for male Speakers. Trained on male speakers

| | birds | jungle | motorcycle | ocean | keyboard | machinegun |
|---|---|---|---|---|---|---|
| IBM | 0.72 | 0.69 | 0.67 | 0.58 | 0.78 | 0.77 |
| IRM | 0.8 | 0.78 | 0.79 | 0.72 | 0.82 | 0.84 |

Table 4.8: Test Result (STOI) for 0 db SNR for male Speakers. Trained on male speakers

| | birds | jungle | motorcycle | ocean | keyboard | machinegun |
|---|---|---|---|---|---|---|
| IBM | 0.83 | 0.8 | 0.77 | 0.71 | 0.85 | 0.83 |
| IRM | 0.88 | 0.87 | 0.85 | 0.82 | 0.87 | 0.9 |

Table 4.9: Test Result (STOI) for 5 db SNR for male Speakers. Trained on male speakers

| | birds | jungle | motorcycle | ocean | keyboard | machinegun |
|---|---|---|---|---|---|---|
| IBM | 0.89 | 0.87 | 0.84 | 0.82 | 0.89 | 0.9 |
| IRM | 0.93 | 0.92 | 0.89 | 0.89 | 0.92 | 0.93 |

Table 4.10: Test Result (STOI) for -5 db SNR for female Speakers. Trained on male speakers

| | birds | jungle | motorcycle | ocean | keyboard | machinegun |
|---|---|---|---|---|---|---|
| IBM | 0.53 | 0.5 | 0.5 | 0.39 | 0.67 | 0.62 |
| IRM | 0.73 | 0.68 | 0.72 | 0.6 | 0.77 | 0.8 |

Table 4.11: Test Result (STOI) for 0 db SNR for female Speakers. Trained on male speakers

| | birds | jungle | motorcycle | ocean | keyboard | machinegun |
|---|---|---|---|---|---|---|
| IBM | 0.63 | 0.58 | 0.59 | 0.51 | 0.7 | 0.67 |
| IRM | 0.82 | 0.78 | 0.8 | 0.73 | 0.82 | 0.85 |

Figure 4.1: STOI is a function of the clean and degraded speech, which are first decomposed into DFT-based, one-third octave bands. Next, short-time (384 ms) temporal envelope segments of the clean and degraded speech are compared by means of a correlation coefficient. Before comparison, the short-time degraded speech temporal envelopes are first normalized and clipped (see text for more details). These short-time intermediate intelligibility measures d(j, m) are then averaged to one scalar value, which is expected to have a monotonic increasing relation with the speech intelligibility.

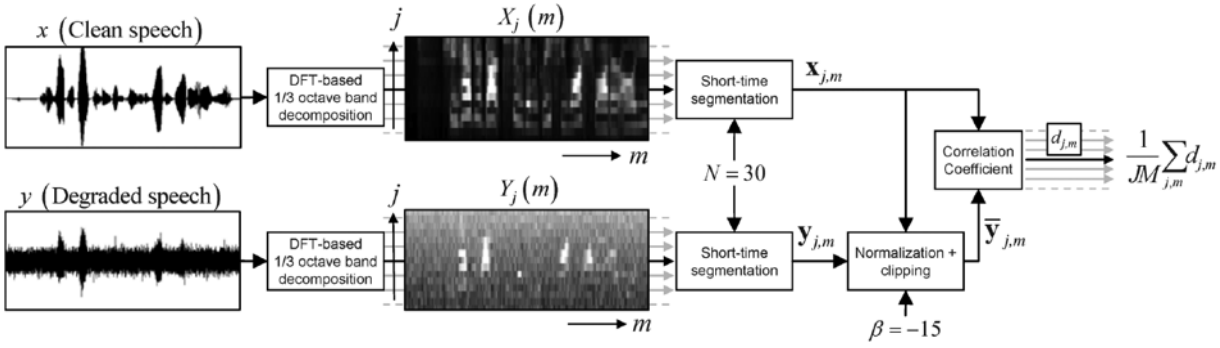


Table 4.12: Test Result (STOI) for 5 db SNR for female Speakers. Trained on male speakers

| | birds | jungle | motorcycle | ocean | keyboard | machinegun |
|---|---|---|---|---|---|---|
| IBM | 0.68 | 0.67 | 0.65 | 0.63 | 0.73 | 0.71 |
| IRM | 0.88 | 0.86 | 0.85 | 0.81 | 0.86 | 0.88 |

from 10 different speech samples. For test data, we used 6 different noises to create noisy speeches. All other data preprocessing steps remains the same.

The basic structure of STOI is illustrated in Fig. 4.1. The output of STOI is a scalar value which is expected to have a monotonic relation with the average intelligibility (e.g., the percentage of correctly understood words averaged across a group of users). A sample-rate of 10 kHz is used, in order to capture a relevant frequency range for speech intelligibility [14]. First, both signals are TF-decomposed in order to obtain a simplified internal representation resembling the transform properties of the auditory system. This is obtained by segmenting both signals into 50% overlapping, Hann-windowed frames with a length of 256 samples, where each frame is zero-padded up to 512 samples.

Before evaluation, silent regions which do not contribute to speech intelligibility are removed. This is done by first finding the frame with maximum energy of the clean speech signal. Both signals are then reconstructed, excluding all the frames where the clean speech energy is lower than 40 dB with respect to this maximum clean speech energy frame. Then, a one-third octave band analysis is performed by grouping DFT-bins. In total 15 one-third octave bands are used, where the lowest center frequency is set equal to 150 Hz and the highest one-third

octave band has a center-frequency equal to approximately 4.3 kHz.

Let x(k,m) denote the k-th DFT-bin of the m-th frame of clean speech. The norm of the j-th one-third octave band, referred to as a TF-unit, is defined as ,
$X_j(m) = \sqrt{\sum_{k=k_1(j)}^{(k_2(j)-1)} x(k,m)^2}$
where, $k_1$ and $k_2$ denote the one-third octave band edges, which are rounded to the nearest DFT-bin. The TF-representation of the processed speech is obtained similarly, and is denoted by $Y_j(m)$.

The model has been trained on all scales on all noises irrespective of voice of genders. But we test our model on different noises separately. All the values in the cell reports STOI explained above ranging from [0,1]. Table 4.1, 4.2 and 4.3 corresponds to male test voice on -5, 0 and 5 db respectively whereas Table 4.4-4.6 shows results on female voices for three different scales. First and second row compares two different targets described in section 3 for comparative analysis. STOI values above 0.6 is considered a good recovery. In that respect, the model performed pretty good for a generalized one. For any particular gender, three scales of SNR values shows recovery for 5db is better than 0db which in turn is better than -5db as expected. Comparing the targets, IBM consistently outperformed IRM on all test data. This coincides with the argument proposed in [4] for the effectiveness of IBM for generalized cases.

Table 4.7-4.12 shows a different experimental setup where we train our model on constrained data. Table 4.7-4.9 shows performance of model trained on male speeches and tested on male speeches too i.e. a matched case. Compared to previous tables result trained on generalized model, the performance is higher which is expected as the model is more adapted the voice. Similarly table 4.10-4.12 shows the unmatched case where female speech is tested with a model trained on male speeches. Expectedly, the performance is lower than generalized model. Also IRM in general performed better than IBM for matched and mismatched case, in contrast to generalized case.

# 5 Future Scopes

An extension of this work has been proposed in [11] where the selection of model has been automated. Proposed modular neural network consists of DNN in lower levels , each trained on a separate noise. Speech auto encoder picks up the best performing DNN at run-time without any interventions. Apart from performing without human intervention, the model also can operate on pre-trained DNNs without any refining.

## 6 CONTRIBUTIONS

Sravan Patibandla

- Configured the google computing cloud setup for processing massive amounts of data. Also, configured the shared services for deploying python, tensorflow on the cloud platform
- Worked on the data preprocessing and building models and signal regeneration
- Contributed toward designing the test scenarios and it?s implementation

Jaideep Patel

- Worked on building the training model using keras and tensorflow.
- Contributed towards the research and development of STOI metric model using matlab and python.
- Contributed towards the testing of matched and unmatched conditions on different noise types, SNR scales and signal types.

Mrinmoy Maity

- Performed research on existing methods and experimental setup
- Implemented the initial skeleton for the model using theano (lasagne). Due to high computational time and configurational limitations on google compute engine, we had to look for other options like keras and tensorflow

## 7 REFERENCES

[1] P.C. Loziou, Speech Enhancement: Theory and Practice. Boca Raton, FL USA: CRC 2007

[2] Wang Y., Narayanan A., Wang D., On Training Targets for Supervised Speech Separation

[3] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. J. Machine Learning Res. 15, 1929?1958 (2014).

[4] D.Wang. On ideal binary mask as the computational goal of auditory scene analysis ?, Speech Separation by Human and Machines. Norwell MA, USA: Kluwer 2005

[5] A. S. Bregman, Auditory scene analysis, Cambridge MA: MIT Press, 1990

[6] A. M. Reddy and B. Raj, Soft mask methods for single-channel speaker separation,? IEEE Trans. Audio, Speech, Lang. Process., 2007.

[7] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In NIPS 2012.

[8] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. Technical report, arXiv:1409.4842, 2014

[9] Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. Fast and robust neural network joint models for statistical machine translation. In Proc. ACL 2014.

[10] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In NIPS 2014.

[11] Minje Kim. Collaborative deep learning for speech enhancement: A run time model selection method using autoencoders

[12] N. Mohammadiha, P. Smaragdis, and A. Leijon, Supervised and un- supervised speech enhancement approaches using nonnegative matrix factorization, IEEE Trans. Audio, Speech, Lang. Process, 2013

[13] R. Hendriks, R. Heusdens, and J. Jensen, MMSE based noise PS tracking with low complexity, in Proc. ICASSP, 2010, pp. 4266?4269.

[14] N. R. French and J. C. Steinberg, ?Factors governing the intelligibility of speech sounds,? J. Acoust. Soc. Amer., vol. 19, no. 1, pp. 90?119,1947.