SPEECH SEPARATION IN SUPERVISED SETTING USING LSTMS

Sravan Patibandla

1. MOTIVATION

Source separation of audio signals is relevant in numerous real world applications. From communication systems to hearing aids, we need to separate noise from signals. Until the dawn of machine learning, we used filter techniques and matrix factorization techniques to solve these problems. Creating filters is a demanding process and we must be able to estimate the properties of each noise we deal with. But as machine learning techniques, especially neural networks have started to prove useful in solving this problem.

The greatest advantage with neural networks is that we can generalize over various noises, given we have massive amounts of data. By generalizing the model for a multitude of noises, we can significantly improve the noise filtering in communication systems, music recording etc. Source separation allows us to remove traffic noise from speech signal during a telephone call or crown noise while recording a concert etc. In the first scenario, we expect minimal latency. However, in the second case, a little latency is acceptable. Hence there is a trade-off between application and latency. In communication systems, the aim is to reduce latency as much as possible. While in systems like recordings, as little latency can be spared.

Previously, several masking techniques have been proposed for speech separation problems as targets for Neural Networks. Ideal Binary Mask (IBM), Ideal Ratio Mask (IRM) are few such masks which have been used as targets. The neural networks explored so far are Deep Neural Networks (DNN) which have many layers of dense units. In this work, we explore the use of Long Short-Term Memory – Recurrent Neural Networks to solve the source separation problem.

2. RELATED WORKS

Human auditory system performs 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 [2], deep neural networks are used to separate noise using various time-frequency masks.

The works stated earlier consider each frame of noise is independent from the previous and following frames. In this work, LSTM-RNNs are explored to capture the time dependency in noise signals. LSTMs look up to 'k' previous frames to identify any patterns in the noisy signal. This paper proposes a 2-layer LSTM-RNN followed by a 4-layer neural network with k=20 and k=40. Section 3 will discuss about LSTMs in general followed by section 4 which will discuss about the basic framework adopted for monaural speech separation. In section 6, the experimental setup and results obtained are presented. Section 5 will conclude our findings and discuss about future scope of the project.

3. LSTM-RNNs

Recurrent Neural Networks (RNNs) are Neural Networks that have the provision to persist information from previous frame to the current and next, so on. We can consider an input of a layer x to be the output of the layer x itself. The following image shows a simple representation of RNNs.

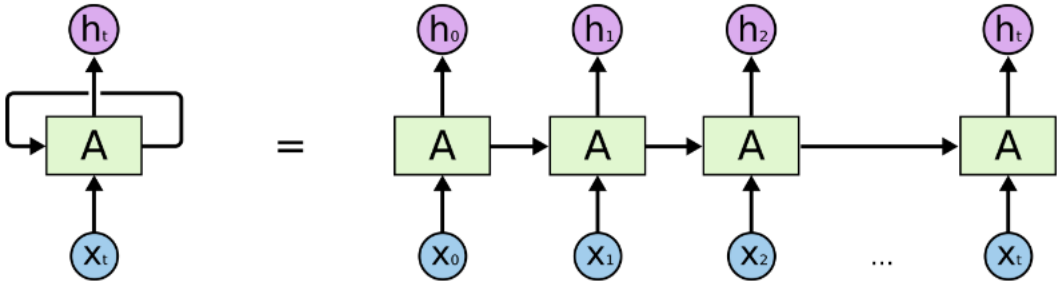

Figure 1: Simple RNN and unrolled version of RNN

To update the model, backpropagation must be done in time too. However, as the input size increases, the problem of vanishing gradients occurs. Due to this, RNNs cannot capture long term dependencies in the input data. To avoid this problem, we use LSTMs, which are an improved version of RNNs.

LSTMs are improved RNNs which can capture long term dependencies, where RNNs failed. LSTMs improve upon the structure of simple RNNs. Instead of information passing through a single layer, LSTMs implement gates to control where the information is flowing. This gives a control over reducing the vanishing gradients problem.

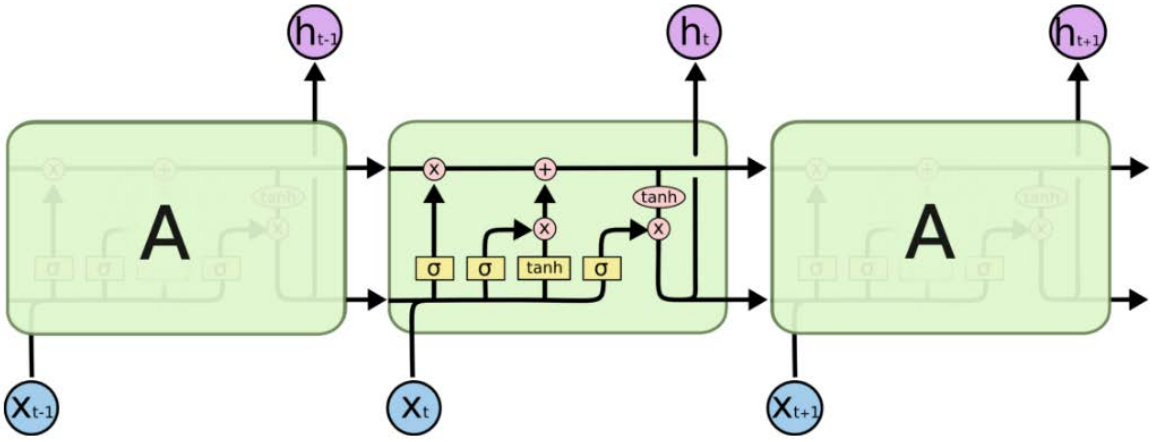

Figure 2: Representation of LSTMs

The upper line is the path of cell state, represented by C_t. The input X_t of current cell and output h_{t-1} of previous layer is passed through a sigmoid layer, which decides what information from previous cell state will remain. This is called the forget layer. Next the input is passed through sigmoid and tanh activation layers. Sigmoid defines what values are to be updated and tanh defines the amount of update. The outputs from these layers are multiplied pointwise and added to the previous cell state. Next, the input is passed to sigmoid layer and the current cell state to tanh layer. The results are multiplied element wise to obtain the output of current cell. This process repeats itself.

4. MODEL FRAMEWORK AND TARGETS

We can visualize our model as a deterministic non-linear mapping function f: N → 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 the experiments to be discussed, amplitudes of short-term Fourier transform 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. However, the results presented are of IBM target. 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. Before feeding data to DNNs, the input is processed by LSTM-RNN layers.

There are various targets which can be used to build the models. However, in this work, we only deal with Ideal Binary Mask (ISM) owing to its simplicity. Other types of targets are Ideal Ratio Masks (IRM), FFT-Mask etc.

5. 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 where the noise signal is superseding speech for a given 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.

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

6. EVALUATION & EXPERIMENTAL RESULTS

The models have been trained on Future Systems at Indiana University. To expedite training, GPUs we employed for training and testing purposes. The data used for the experiments is from TIMIT dataset, which has 8 different dialects and male and female speakers of all the dialects. The models

are trained using data corrupted using 5 different noises, namely birds, casino, jungle, motorcycle, ocean at different SNR levels (0dB, -5dB, 5dB). The test data contains samples corrupted with the noises present in training data and an additional noise called *computerkeyboard*.

There are two architectures which have been explored. The first and second architecture have two LSTM layers stacked together, followed by 3 hidden layers and an output layer. The difference between first and second architectures is in the lookback of LSTMs. The first model has a lookback of 40 frames while the second has a lookback of 20 frames. The hidden layers have a dropout rate of 0.2. The models are learnt using *adam* optimizer which has been proved to be a superior optimizer for deep neural networks and LSTMs as well.

The test and training data contain 150 randomly sampled speech samples for each noise type. The same sample for a given noise type is used to produce all the scales of noise corruption. The time-frequency spectrum is obtained by applying short-term Fourier transform data with frame size 1024 samples and overlap of 512 samples and hann window. For evaluation, we use STOI score which ranges from 0 to 1, 1 being the best. The other scale for evaluation is Signal to Distortion Ratio (SDR), which measures the distortion in the signal.

Using the 2 different architectures, a total of six models are built. Each architecture is fed with three different types of dataset. One dataset comprises of both male and female samples, the second one being exclusively male samples, and the third being exclusively female. This helps us in evaluating the matched and unmatched performance in all these models.

Table 1: Architecture 1 trained with both male and female speech samples (lookback = 40 frames)

		Noises											
		birds		casino		jungle		motorcycle		ocean		computerkeyboard	
		SDR	STOI	SDR	STOI	SDR	STOI	SDR	STOI	SDR	STOI	SDR	STOI
Male	-5 dB	4.6	0.53	-3.4	0.45	-3.7	0.52	1.17	0.55	-2.4	0.49	-2.7	0.5
	0 dB	7	0.63	1.6	0.55	1.4	0.6	5	0.62	2.3	0.57	1.75	0.58
	5 dB	8.8	0.69	5.6	0.62	5.8	0.66	7.6	0.67	6.1	0.64	5.4	0.65
Female	-5 dB	3.7	0.46	-3.8	0.4	-3.7	0.48	0.9	0.52	-2.6	0.45	-2.76	0.46
	0 dB	6.4	0.57	1.35	0.51	1.26	0.56	4.7	0.58	2.3	0.53	1.6	0.53
	5 dB	8.0	0.63	5.5	0.58	5.7	0.62	7.3	0.62	6	0.6	5.2	0.59

Table 2: Architecture 1 trained with male speech samples (lookback = 40 frames)

		Noises											
		birds		casino		jungle		motorcycle		ocean		computerkeyboard	
		SDR	STOI	SDR	STOI	SDR	STOI	SDR	STOI	SDR	STOI	SDR	STOI
Male	-5 dB	3	0.53	-0.9	0.39	0.54	0.48	3.4	0.56	-1.3	0.4	-4.2	0.55
	0 dB	5.7	0.63	2.3	0.5	3.2	0.57	6.4	0.66	2.53	0.5	0.75	0.65
	5 dB	8.13	0.72	4.8	0.6	5.7	0.66	8.8	0.73	5.77	0.63	5.17	0.73
Female	-5 dB	1	0.39	-2.5	0.27	-0.7	0.37	2.5	0.46	-2.7	0.27	-5	0.47
	0 dB	3.75	0.49	0.57	0.36	1.73	0.44	5.3	0.55	1.24	0.36	0.05	0.56
	5 dB	6.1	0.59	3.12	0.46	3.86	0.51	7.4	0.63	4.17	0.49	4.33	0.65

Table 3: Architecture 1 trained with female speech samples (lookback = 40 frames)

		Noises											
		birds		casino		jungle		motorcycle		ocean		computerkeyboard	
		SDR	STOI	SDR	STOI	SDR	STOI	SDR	STOI	SDR	STOI	SDR	STOI
Male	-5 dB	-0.2	0.35	-3.3	0.26	-3	0.32	1.45	0.46	-1.3	0.3	-2.15	0.43
	0 dB	1.9	0.44	-0.2	0.35	0.12	0.4	3.4	0.54	1.3	0.4	-0.8	0.51
	5 dB	3.7	0.53	1.67	0.43	3.44	0.47	5	0.6	3.26	0.5	2.7	0.57
Female	-5 dB	3.1	0.45	0.13	0.34	0.2	0.38	3.1	0.45	1.6	0.37	-3.4	0.49
	0 dB	5.0	0.54	3.02	0.44	3.23	0.47	6.77	0.61	4.23	0.48	1.2	0.57
	5 dB	6.84	0.62	5.2	0.53	5.6	0.56	8.65	0.68	6.67	0.59	5.3	0.64

Presented above are the results of three models built with the first architecture: 2 stacked LSTM layers with lookback of 40 frames, followed by 3 dense layers and an output layer. Table 1, Table 2, and Table 3 are the results of test data of first architecture trained on both gender speech samples, trained on exclusively male samples, and trained on exclusively female samples respectively.

Table 4: Architecture 2 trained with both male and female speech samples (lookback = 20 frames)

		Noises											
		birds		casino		jungle		motorcycle		ocean		computerkeyboard	
		SDR	STOI	SDR	STOI	SDR	STOI	SDR	STOI	SDR	STOI	SDR	STOI
Male	-5 dB	6	.59	-1.3	0.38	-1.1	0.48	2.13	0.58	1.04	0.47	-3	0.56
	0 dB	8.7	0.69	2.67	0.53	3	0.6	5.8	0.69	4.48	0.59	1.6	0.66
	5 dB	11.1	0.77	6	0.66	6.55	0.7	9.2	0.77	7.7	0.71	6.2	0.76
Female	-5 dB	4.87	0.52	-1.4	0.34	-1.2	0.43	1.96	0.55	0.86	0.4	-3.3	0.53
	0 dB	7.6	0.64	2.6	0.49	2.9	0.55	5.87	0.66	4.35	0.54	1.4	0.63
	5 dB	9.9	0.73	5.8	0.6	6.3	0.66	9.2	0.75	7.48	0.66	5.9	0.73

Table 5: Architecture 2 trained with male speech samples (lookback = 20 frames)

		Noises											
		birds		casino		jungle		motorcycle		ocean		computerkeyboard	
		SDR	STOI	SDR	STOI	SDR	STOI	SDR	STOI	SDR	STOI	SDR	STOI
Male	-5 dB	2.15	0.53	-1.1	0.41	1.5	0.56	-5	0.5	-5	0.39	-3.9	0.54
	0 dB	5.3	0.63	2.6	0.54	4.3	0.64	0.09	0.59	-0.3	0.5	0.56	0.63
	5 dB	7.7	0.72	5.23	0.65	6.45	0.7	4.3	0.68	3.68	0.61	4.87	0.73
Female	-5 dB	0.2	0.44	-2.4	0.34	0.42	0.47	-6.8	0.42	-6.8	0.31	-5.4	0.45
	0 dB	3.2	0.54	1.11	0.44	2.88	0.55	-1.53	0.5	-2	0.4	-0.76	0.55
	5 dB	5.4	0.63	3.4	0.55	4.6	0.62	2.6	0.59	2.77	0.51	3.38	0.65

Table 6: Architecture 2 trained with female speech samples (lookback = 20 frames)

		Noises											
		birds		casino		jungle		motorcycle		ocean		computerkeyboard	
		SDR	STOI	SDR	STOI	SDR	STOI	SDR	STOI	SDR	STOI	SDR	STOI
Male	-5 dB	-0.2	0.36	-3.9	0.24	-3.6	0.3	1.4	0.43	-1.6	0.29	-5.1	0.45
	0 dB	1.9	0.44	-0.9	0.32	-0.2	0.38	3.35	0.5	0.96	0.38	-0.86	0.51
	5 dB	3.67	0.51	1.06	0.41	2.1	0.45	4.7	0.6	2.8	0.48	2.52	0.56
Female	-5 dB	3.02	0.45	0.04	0.34	0.08	0.38	4.3	0.52	1.45	0.37	-3.53	0.5
	0 dB	5.1	0.55	2.9	0.45	2.2	0.48	6.8	0.6	4.2	0.48	1.1	0.6
	5 dB	6.83	0.63	5.2	0.55	5.7	0.57	8.7	0.68	6.6	0.6	5.14	0.66

Presented above are the results of three models built with the second architecture: 2 stacked LSTM layers with lookback of 20 frames, followed by 3 dense layers and an output layer. Table 4, Table 5, and Table 6 are the results of test data of first architecture trained on both gender speech samples, trained on exclusively male samples, and trained on exclusively female samples respectively.

The *computerkeyboard* noise is unseen to the models, which are trained on *birds*, *casino*, *jungle*, *motorcycle*, and *ocean* noises. The striking observation is the models can generalize on unseen noises and have a comparable performance with seen noises. The models perform much better when a higher SNR signal is tested. The performance drastically reduces as the strength of the noise increases.

In unmatched cases, i.e. when the model is trained with exclusively one gender data, we can observe that the test performance on other gender is worse. However, the performance of matched voices is comparable to the performance of models trained with both speech samples. It

can be extrapolated that ensemble of male speech trained model and female speech trained model perform no better than model trained with both gender's samples, if not worse.

7. CONCLUSION & FUTURE SCOPE

The models perform well on unseen noises and the performance of models trained with speech samples of both the genders is much better than any model trained exclusively with one gender.

My observation in using LSTMs is that they take a much longer time to train and there is significant latency in test phase too. In my previous work, I used Deep Neural Networks of comparable size and same train data. On a meager 2 GPUs, the training time for 1 epoch was around 2 seconds. But, with LSTMs, the training time has sprung to ~30 minutes approximately. This could be a major hindrance in adoption of LSTMs for source separation problems.

Due to time constrained I could not present the results of another architecture where 3 stacked LSTMs are used. The training time for 1 epoch was about 45 minutes. After 20 iterations, I could not regenerate a decent signal. Efficient structure of LSTMs might decrease the training duration.

8. 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 andMachines. NorwellMA, USA: Kluwer 2005

[5] A. S. Bregman, Auditory scene analysis, CambridgeMA: 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 JohnMakhoul. 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] 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

[12] 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.

[13] Lei Sun, Jun Du et. Al., Multiple-target Deep Learning for LSTM-RNN BASED Speech Enhancement, IEEE HSCMA 2017

[14] http://colah.github.io/posts/2015-08-Understanding-LSTMs/