Artificial Intelligence for Human-Robot Interaction
AAAI Technical Report FS-17-01

Who Said That? A Comparative Study of Non-Negative Matrix Factorisation and Deep Learning Techniques

Teun F. Krikke, Frank Broz, David Lane
Heriot-Watt University
Edinburgh Campus
Riccarton, UK EH14 4AS

Abstract

When working with robots it is very important that the robot understands the user. This is more difficult when the user is only able to speak to it. You do not want a robot to call for milk when the user said call for help. It is possible for a robot to get a clear understanding of the user in a lab environment where there is no noise or reverberation to distort the instructions. However, in a normal setting this is not always the case. We concentrate on speaker separation to improve speech recognition. To do this we use non-negative matrix factorisation (NMF) and deep learning techniques. For training and testing these techniques, we introduce a *new corpus* that is recorded with a microphone array. In this paper, we use different NMF and deep learning techniques for the speaker separation. We found that adding directional information improves the separation when there is no noise or reverberation. However, when reverberation is present we saw that the NMF technique with the Itakura-Saito cost function out performs the other techniques. With deep learning we found that a recurrent neural networks is able to perform the separation of the speakers.

Introduction

With the introduction of devices, such as Google Home and Amazon Echo, the importance of speech recognition has increased. However, speech recognition still assumes a clean environment (no noise or reverberation) and finds it difficult to understand people in real-life environments. In order to work correctly, devices like Google Home assume that the speaker can be clearly understood. To help with this we use deep learning (Huang et al. 2014; Kang et al. 2015; Nugraha, Liutkus, and Vincent 2016; Weninger et al. 2014) and non-negative matrix factorisation (NMF) (Févotte, Bertin, and Durrieu 2009; Grais and Erdogan 2011; Parathai et al. 2015; Stein 2014) to separate the speaker from background noise or other speakers. For NMF, the choice of cost function is very important. The right cost function can improve the performance of NMF significantly.

We use four different deep learning networks and three different NMF cost functions which are trained and tested on three corpora. The three corpora we use are an acoustic-camera (AC) corpus, a vocalization corpus (Salamin, Polychroniou, and Vinciarelli 2013) and the map task corpus (Anderson et al. 1991). The AC corpus is the only one that contains noise and is used to test the performance of the algorithm when there is additional noise and reverberation in the recordings. For deep learning techniques, we use a long short-term memory (LSTM) network, convolution neural network (CNN), deep neural network (DNN) and a recurrent convolution neural network (RCNN). As cost functions for NMF, we choose Itakura-Saito (Févotte, Bertin, and Durrieu 2009), Euclidean and Kullback-Leibler.

The novelty of this paper mainly in the usage of the acoustic-camera corpus that we created. This corpus contains far-field recordings of two speakers speaking with overlapping speech. The speakers are moving through the room whilst speaking. Therefore the corpus can also be used for speaker tracking. The microphone array (which is called the acoustic-camera) is able to make far field recordings (>5 metres away from the AC) and locate the sound sources in these recordings.

Corpus

We use three different corpora for testing the NMF and deep learning techniques. A concise overview of the three corpora is given in Table 1.

The first corpus is a small corpus recorded with the AC (see Table 1). This device contains 72 microphones, which are placed in a circular configuration with a diameter of 1 metre and one camera in the middle of the circle. The AC is capable of recording for a maximum of 1:30 minutes at a frame rate of 192kHz for the audio recordings. It gives us an exact location of the microphones and allows us to use beam-forming to get an approximate location of the sources. Beam-forming uses the recordings from all microphones to determine which direction the sound is coming from (Döbler and Heilmann 2008; Schröder and Jaeckel 2012). In a clean environment, the AC is able to locate the origin of the sound (see Figure 1), but with multiple sound sources, it is not able to separate them. The room used for recordings has noise from the air-conditioning along with reverberation due to the room size, as is typical of many home and office environments. The high sensitivity of the microphones to noise and echo means that post processing is needed to create a clear approximation of the source location.

We made 7 recordings with the AC. Of these 7 recordings 5 are female-male and 2 are male-male recordings. In

corpus	# subjects	# mics	# files	file length	separated ground truth	noise	distance microphone to source	recording environment	F_s	transcripts
vocalization corpus	120 (63 women, 57 men)	1 (per file)	2763	0:10	Yes	No	$<$ 1 metre	Lab setting	16kHz	No
map task corpus	64	1 (per file)	191	5:00	No	No	$<$ 1 Metre	Lab setting	20kHz	Yes
acoustic-camera	9 (6 men, 3 women)	72 (per file)	7	1:30	No	Yes	$>$ 6 metres	Empty room	192kHz	Yes

Table 1: Overview of the different corpora.

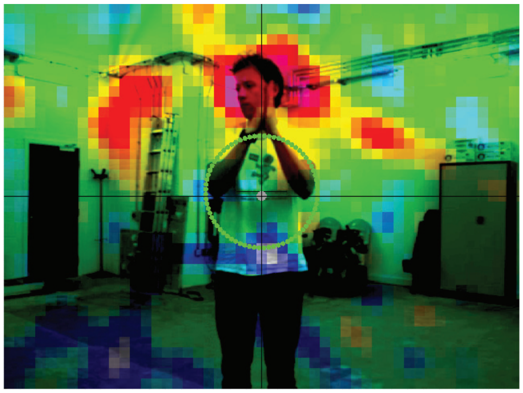

Figure 1: Pressure map showing the origin of the sound (pink blob) and reverberation (red blobs)

total we have 9 different speakers, 3 women and 6 men. All speakers were given the first 7 pages out of the book A way in the wilderness by R. M. Ballantyne[1] to read aloud. Therefore, the speech can easily be transcribed. The speakers were instructed to stand still for two thirds of the recorded time, after which they should walk around the room keeping a minimal distance of 6 metres away from the camera. Each of these seven recordings contains two speakers and were made in a room of 9 by 13 metres, in total we have recorded 9 different speakers. The AC does not provide us with depth information but does give us a video recording of the speaker. This, we can use for visual tracking.

The second corpus that we use is the vocalization corpus[2] (Salamin, Polychroniou, and Vinciarelli 2013) which contains recorded telephone conversations of 120 different subjects. In the recording there is background speech present of a second speaker. This does not provide us with a clean ground truth. For this corpus there is no localisation information available.

The map task corpus (Anderson et al. 1991) is the third corpus we are using. In this corpus people have headphones and a microphone and need to explain to each other how to get from A to B on a map. This corpus contains speech of 64 subjects. In the recordings the second speaker can be heard in the background. This means that this does not provide us with a clean ground truth nor does this corpus have localisation information available.

For the corpora that do not have localisation information, we have created this artificially when it is needed. To do this we use a time delay of one audio frame. This means that our artificially created microphones spaced at a relative distance of 1 audio frame apart. This is dependent on the frame rate of the recording. For example, when a recording is made at 16 kHz the microphones would be spaced at $\frac{speedofsound}{framerate}$ metres or in this case $\frac{340.29}{16000}$ metres which is equal to 0.021 metres.

Techniques

As input data for the different techniques we are using the short-time Fourier transform (STFT) with a window of 30 ms and an overlap of 10 ms. The window size is slightly bigger than what is normally used in speech applications (25 ms) and should pick up speech better than shorter windows. The amount of overlap is the same as what is normally used for speech recognition. We use the vocalization corpus and map task corpus to determine how well each technique performs the separation task. With our own corpus we measure the performance of the different techniques when there is noise and reverberation in the recording. To deal with this, we applied some preprocessing techniques in the form of noise reduction and a multi-band compressor for reverberation reduction. This gave us four different sets of files; one without both reverberation and noise, one with only reverberation, one with only noise and the original recording containing both.

Non-negative matrix factorisation

NMF is a clustering algorithm that tries to approximate the input signal (X) which is the squared magnitude information of the recordings. To ensure that the input to the algorithm is non-negative we square the outcome of the STFT. It does the approximation by multiplying two matrices (W and H) together (see Equation 1). The W matrix is an approximation of the signal coming from the different sources (K) and

[1] http://www.gutenberg.org/ebooks/21715?msg=

[2] http://www.dcs.gla.ac.uk/vincia/?p=378

Technique	cost function	parameters
Sparse Euclidean	Euclidean	$\lambda = 0.0001$
Convolution Euclidean	Euclidean	
IS	Itakura-Saito	
Sparse IS	Itakura-Saito	$\lambda = 0.0001$
Convolution IS	Itakura-Saito	
Sparse KL	Kullback-Leibler	$\lambda = 0.0001$
Convolution KL	Kullback-Leibler	
DoA	Kullback-Leibler	
TDoA	Kullback-Leibler	

Table 2: Overview of the different NMF techniques and cost functions.

the H matrix is an approximation of the gain of the different sources. When multiplied together, this gives us the approximated version of X ($\widetilde{X}$). Assuming that the size of X is frequency (F) multiplied by time (N) then the size of the matrix W is F x K and the size of H is K x N. The difference between the approximated version and the input signal is the cost (see Equation 2) which can be calculated using different cost functions. In our case, we are using Kullbeck-Leibler divergence (see Equation 4), Euclidean distance (see Equation 3) and Itakura-Saito divergence (see Equation 5) as cost functions. Apart from the different cost functions we use four different versions of NMF (see Table 2) these are: direction of arrival, time-difference of arrival, sparse and convolution. (Stein 2014) describes the direction of arrival NMF in an additional version called non-negative tensor factorisation where the direction of arrival is used as an extra dimension. For completeness we have also applied this to our corpora. Two of the techniques require extra information in the form of direction of arrival (Stein 2014) or time difference of arrival (Nikunen and Virtanen 2014) to separate the sources.

$$X \approx \widetilde{X} = WH \tag{1}$$

$$D(X|\widetilde{X}) = \sum_{f=1}^{F} \sum_{n=1}^{N} d([X]_{fn}|[\widetilde{X}]_{fn}) \tag{2}$$

$$d_{EUC}(x|y) = \frac{1}{2}(x-y)^2 \tag{3}$$

$$d_{KL}(x|y) = xlog\frac{x}{y} - x + y \tag{4}$$

$$d_{IS}(x|y) = \frac{x}{y} - log\frac{x}{y} - 1 \tag{5}$$

Deep learning

As input for the Deep learning techniques we use the unmodified STFT of the input signal. We use 4 different networks:

- recurrent neural network (RNN) with two long-short term memory (LSTM) nodes with 512 units (Huang et al. 2014)
- convolution neural network (CNN) with two convolution layers with 64 filters of 3 x 3 (Choi, Fazekas, and Sandler)
- deep neural network (DNN) with two hidden layers (Huang et al. 2014) each with 150 units
- recurrent convolution network (RCNN) with two convolution layers (64 filters of 3 x 3) and one recurrent layer (512 units)

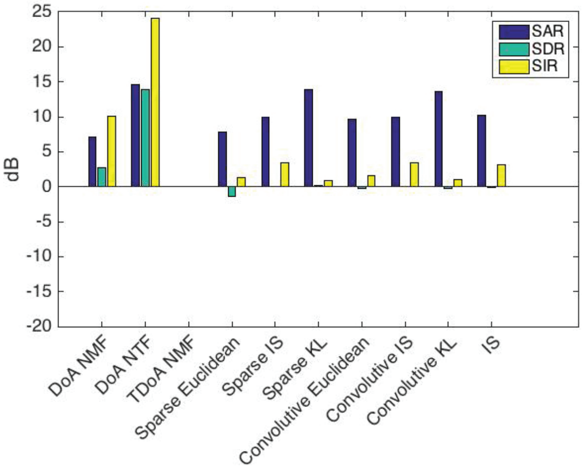

Figure 2: A comparison between different NTF and NMF techniques on the vocalization corpus.

In addition to the current configuration each network has a separation layer which contains an additional hidden layer with 512 units and a Wiener filter that is used for the separation. The loss function we use for optimising the network is the mean squared error between the output of the network, which is two signals, and the ground truth of the sources.

Results

For evaluating the different techniques we apply 3 objective measurements introduced in (Vincent, Gribonval, and Févotte 2006) namely: signal-to-distortion ratio (SDR); signal-to-interference ratio (SIR) and signal-to-artefact ratio (SAR). These measurements take a ground truth and a the outcome of the algorithms for comparison. Positive values indicate better performance for all measurements. If we take for example SIR then a negative value for SIR would mean that there is more information present from the interfering signal than from the ground truth signal.

Looking at the results from applying NMF to the vocalization corpus (see Figure 2), we see that adding directional information gives the algorithm an advantage over the techniques that lack this information. However, when this algorithm is provided with noisy data where the location of the microphones is exact instead of relative, the algorithm preforms worse than much simpler techniques (see Figure 3). This could be because of the distance between the microphones, which is not only exact but also bigger ((Stein 2014) assumes a distance of 1 audio frame between the microphones). The exact distance has a disadvantage because now an audio frame can be lost because of the distance measurement.

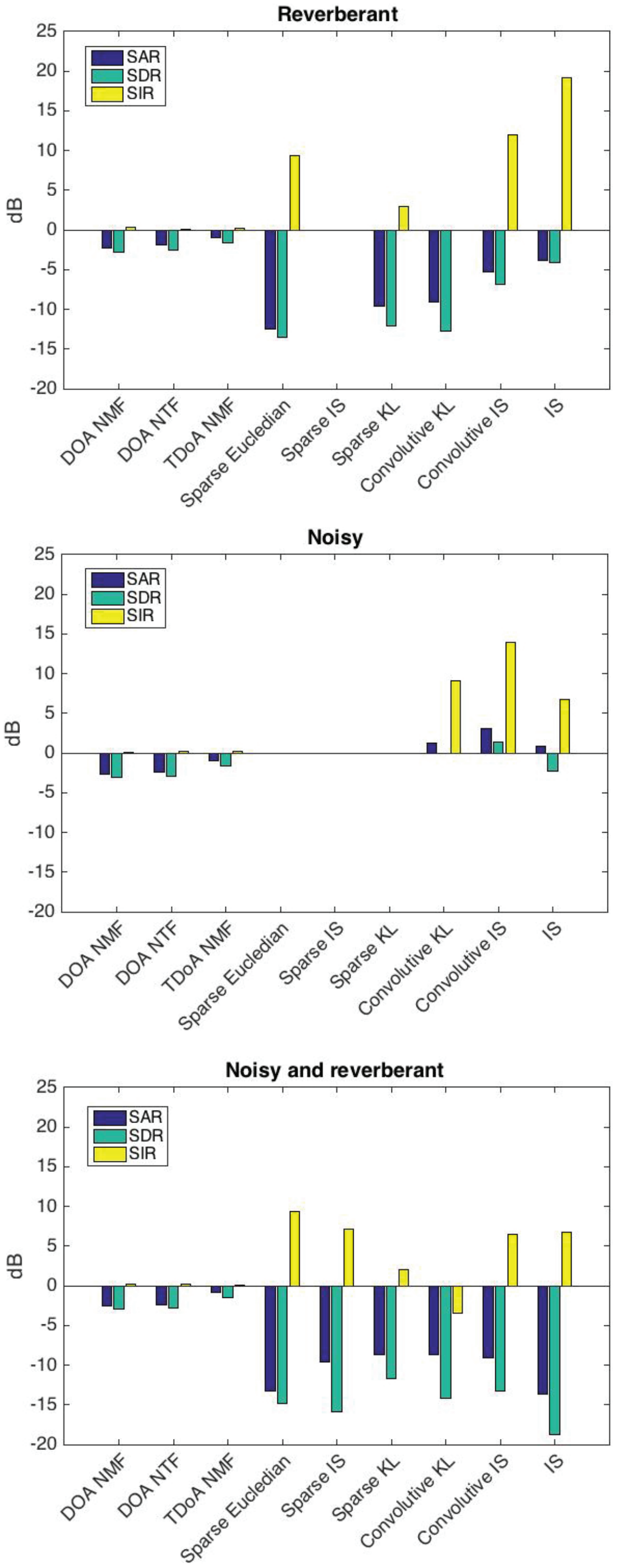

Figure 3: A comparison between different NMF techniques on the reverberant (top), noisy (middle) and original (bottom) recordings from the AC corpus using the cleaned (no noise or reverb) recordings as the ground-truth.

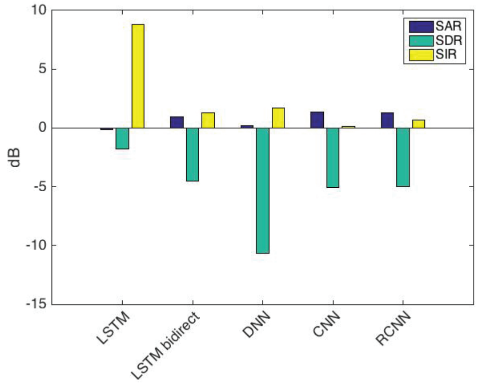

Figure 4: A comparison between different deep learning techniques on the vocalization corpus.

Looking at the deep learning techniques (see Figure 4), we see that the normal LSTM (thus the recurrent network with two LSTM nodes) out performs all techniques. Also, changing this particular network into a forward-backward network where it has information from the past and the future performs worse than the normal LSTM but not worse than the other techniques.

Conclusion

In this paper we applied different techniques to the problem of speaker-speaker separation. As input data to this we used corpora that contain recordings of single speakers as well as our own corpus which contains recordings of two speakers. We have ensured that the input data contained overlapping speech which makes it more difficult for an algorithm to distinguish between speakers. At the time of writing not all experiments have finished. We are still waiting for the results of the speech recogniser, map task corpus and AC corpus.

We found that adding directional information to NMF gives us the best result (as seen by (Stein 2014)) on the vocalization corpora (see Figure 2). However, when there is noise or echo present in the recording, as it the case with the AC corpus, then the Itakura-Saito cost function performs best (3). This cost function is only surpassed by the sparse Euclidean distance when both noise and echo are present. For deep learning we found that a single layer long short-term memory (LSTM) gives the best result. This network out performs a forward-backward 2-layer LSTM, a deep neural network, a convolution neural network and a recurrent convolution neural network (see Figure 4). We also see that NMF out performs deep learning. Unfortunately, NMF does not allow for real-time separation thus can only be used offline.

Comparing the results to the related work we see that our results are worse. The reasons for this is that our corpus contains noise and overlapping speech which makes it harder for an algorithm to separate the speakers then when there is some kind of turn-taking happening. As with the vocalization corpus this was never designed for speaker separation but for laughter detection.

Acknowledgements

This research is supported by the Engineering and Physical Sciences Research Council (EPSRC), as part of the CDT in Robotics and Autonomous Systems at Heriot-Watt University and The University of Edinburgh. Grant reference EP/L016834/1

References

Anderson, A. H.; Bader, M.; Bard, E. G.; Boyle, E.; Doherty, G.; Garrod, S.; Isard, S.; Kowtko, J.; McAllister, J.; Miller, J.; Sotillo, C.; Thompson, H. S.; and Weinert, R. 1991. The hcrc map task corpus. *Language and Speech* 34(4):351–366.

Choi, K.; Fazekas, G.; and Sandler, M. Explaining deep convolutional neural networks on music classification.

Döbler, D., and Heilmann, G. 2008. Time-domain beamforming using zero-padding. In *Berlin Beamforming Conference (BeBeC)*.

Févotte, C.; Bertin, N.; and Durrieu, J.-L. 2009. Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural computation* 21(3):793–830.

Grais, E. M., and Erdogan, H. 2011. Single channel speech music separation using nonnegative matrix factorization and spectral masks. In *17th International Conference on Digital Signal Processing (DSP), 2011.*, 1–6. IEEE.

Huang, P.-S.; Kim, M.; Hasegawa-Johnson, M.; and Smaragdis, P. 2014. Deep learning for monaural speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014.*, 1562–1566. IEEE.

Kang, T. G.; Kwon, K.; Shin, J. W.; and Kim, N. S. 2015. NMF-based target source separation using deep neural network. *IEEE Signal Processing Letters* 22(2):229–233.

Nikunen, J., and Virtanen, T. 2014. Direction of arrival based spatial covariance model for blind sound source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22(3):727–739.

Nugraha, A. A.; Liutkus, A.; and Vincent, E. 2016. Multichannel audio source separation with deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24(9):1652–1664.

Parathai, P.; Woo, W. L.; Dlay, S.; and Gao, B. 2015. Single-channel blind separation using l 1-sparse complex non-negative matrix factorization for acoustic signals. *The Journal of the Acoustical Society of America* 137(1):EL124–EL129.

Salamin, H.; Polychroniou, A.; and Vinciarelli, A. 2013. Automatic detection of laughter and fillers in spontaneous mobile phone conversations. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2013.*, 4282–4287. IEEE.

Schröder, R., and Jaeckel, O. 2012. Evaluation of beamforming systems. In *Proceedings of the 4th Berlin Beamforming Conference*, 22–23.

Stein, N. D. 2014. Nonnegative tensor factorization for directional blind audio source separation. *stat* 1050:18.

Vincent, E.; Gribonval, R.; and Févotte, C. 2006. Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing* 14(4):1462–1469.

Weninger, F.; Hershey, J. R.; Le Roux, J.; and Schuller, B. 2014. Discriminatively trained recurrent neural networks for single-channel speech separation. In *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*, 577–581. IEEE.