Binaural and Log-Power Spectra Features with Deep Neural Networks for Speech-Noise Separation

Alfredo Zermini[1], Qingju Liu[1], Yong Xu[1], Mark D. Plumbley[1], Dave Betts[2], Wenwu Wang[1],
[1]Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, Surrey, GU2 7XH, UK
[2]CEDAR Audio Ltd, 20 Home End, Fulbourn, Cambridge, CB21 5BS, UK
{a.zermini, q.liu, yong.xu, m.plumbley, w.wang}@surrey.ac.uk, dave.betts@cedaraudio.com

***Abstract*—Binaural features of interaural level difference and interaural phase difference have proved to be very effective in training deep neural networks (DNNs), to generate time-frequency masks for target speech extraction in speech-speech mixtures. However, effectiveness of binaural features is reduced in more common speech-noise scenarios, since the noise may over-shadow the speech in adverse conditions. In addition, the reverberation also decreases the sparsity of binaural features and therefore adds difficulties to the separation task. To address the above limitations, we highlight the spectral difference between speech and noise spectra and incorporate the log-power spectra features to extend the DNN input. Tested on two different reverberant rooms at different signal to noise ratios (SNR), our proposed method shows advantages over the baseline method using only binaural features in terms of signal to distortion ratio (SDR) and Short-Time Perceptual Intelligibility (STOI).**

I. Introduction

Source separation is a well-studied topic, with many existing methods available such as independent component analysis [1], computational auditory scene analysis [2], [3], and non-negative matrix factorization [4]. More recently, Deep Neural Networks (DNNs) [5] have shown the state-of-the-art performance in source separation [6]–[8].

In this paper, we focus on the problem of source separation from binaural recordings to mimic human listening, where a target speech signal is embedded by interfering background noise. The method proposed in [8], [9] extracts a target speech signal from a competing speech signal, using a DNN trained using binaural spatial cues of mixing vectors (MV), interaural level difference (ILD) and interaural phase difference (IPD). However, the above spatial cues become less effective for speech-noise scenarios where the target speech is often masked over by the background noise in adverse conditions. Moreover, if the environment is not anechoic, e.g. a reverberant room, performance of the existing method degrades radically.

To address the above limitations, extra information should be exploited. Fortunately, speech sound and background environment noise often bear very essential spectral difference by nature, such as their spectral patterns. For instance, log-power spectra (LPS) have been proved useful as DNN input to extract target speeches corrupted by noise from monaural recordings [10], [11], which provide complementary information to the spatial cues. Therefore, here we explore the use of both binaural features and LPS in the DNN based source separation. We employed a similar DNN structure of softmax classifier, which is performed in several frequency bands independently, instead of the feed-forward regression model as used in [10], [11]. Moreover, we propose to solve the separation problem in more realistic environments such as reverberant rooms, instead of separation from direct summation of speech and noise [10], [11].

The remainder of the paper is organized as follows. Section II introduces the proposed system, including the overall DNN architecture employed, the low-level feature extraction for the DNN input and output in the training stage, the system implementation and evaluation. Experimental results are presented in Section III, where evaluations are performed and analyses are given, followed by conclusions of our findings and insights for future work in Section IV.

II. Proposed Method

A. System overview

The system shown in Figure 1, merges together the information from several DNNs in order to get a series of soft-masks, which are used to separate the speech source from the audio mixture. Figure 1 shows how the system of DNNs works. The short-time Fourier transform (STFT) on the left and right channels is calculated, in order to get the two

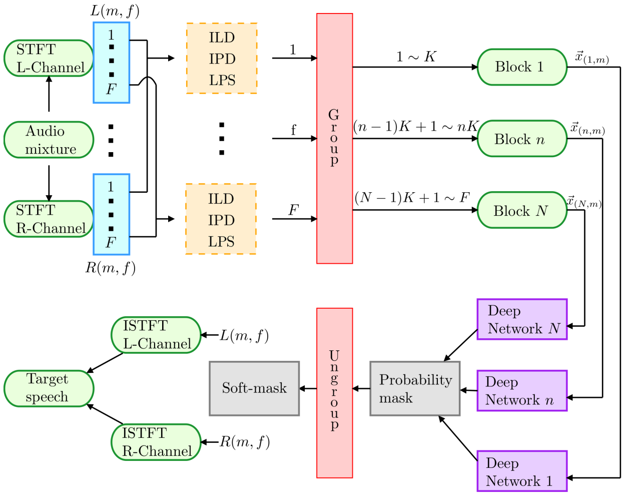


Fig. 1: Diagram of the system architecture using deep neural networks.

978-1-5090-3649-3/17/$31.00 ©2017 IEEE

spectrograms $X_L(m,f)$ and $X_R(m,f)$, where $m = 1, \cdots, M$ and $f = 1, \cdots, F$ are the time frame and frequency bin indices respectively.

For each T-F bin, the low-level features (i.e. ILD, IPD, LPS) are calculated. They will be explained in details in section II-B. The low-level features are arranged into N blocks, each one containing the information for a group of frequency bins. Each of the N blocks, labelled n, includes $K = 8$ frequency bins $((n-1)K+1, \cdots, nK)$, where $K = F/N$ and N is the number of DNNs. In this way, each block contains the information from a very specific set of frequencies. Each block is used as the input of a different DNN, whose output is a softmax classifier containing J values between 0 and 1, each one associated with the probability of a certain Direction Of Arrival (DOA).

After merging all the outputs, a probability mask like the one in Figure 2 can be obtained. As explained in section II-D, a series of soft-masks can be generated from the probability mask, one for each test set j. The soft-masks are multiplied element-wise by the mixture spectrograms and, after applying the inverse STFT (ISTFT), the target source can be recovered, as well as the interferer source.

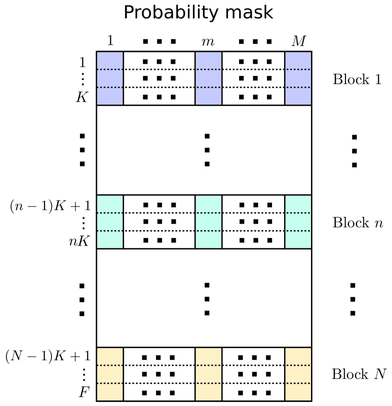


Fig. 2: The probability mask obtained from the output of the DNNs system.

B. Low-level features

The low-level features used are the IPD, ILD and LPS [12] [10]. They are used to derive high-level features which are easy to classify. Unlike [8], [9], the MV feature is omitted since it leads to negligible improvements. IPD and ILD are the phase and the amplitude difference between the left and the right channels and they are respectively given by [13]

$$ILD(m,f) = 20\log_{10}\left(\left|\frac{X_L(m,f)}{X_R(m,f)}\right|\right),$$

$$IPD(m,f) = \angle\left(\frac{X_L(m,f)}{X_R(m,f)}\right).$$

The LPS is defined by

$$LPS(m,f) = \frac{1}{2}\left(\log\left(\left|X_L(m,f)^2\right|\right) + \log\left(\left|X_R(m,f)^2\right|\right)\right).$$

By putting together the ILD, IPD and LPS vectors, one can obtain, for each T-F unit

$$\vec{x}(m,f) = \left[ILD(m,f), IPD(m,f), LPS(m,f)\right]^T.$$

Each $\tilde{u}(m,f)$ is grouped into N blocks along the frequency bins, which represents the input vector of each DNN

$$\vec{x}_{(n,m)} = \left[\vec{x}^T(m,(n-1)K+1), \cdots, \vec{x}^T(m,nK)\right]^T.$$

C. The DNNs structure

Figure 3 shows the structure of DNNs used.

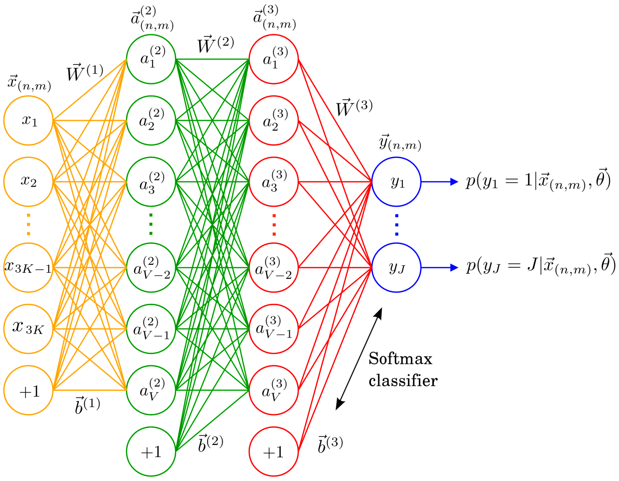


Fig. 3: Structure of a DNN.

D. Soft-masks generation

The output of each DNN looks like the one shown in Figure 4, which is the case where the target and the interferer are located respectively at 0° and −70°, where the SNR between speech and noise is 20 dB. The output represents the DOA estimation, for a given group of frequency bins. By averaging on the time frames belonging to the same test set, a vector with J different values can be obtained. For a given test set j, the T-F bins which belongs to the maximum DOA row of the probability mask are copied to the target mask, while those corresponding to the second highest value are copied to the interferer mask. In the case in Figure 4, the most probable rows correspond to the DOAs labelled 0° and −70°, where the speech and the noise are expected to be found. There are several misplaced bins, which lead to a loss of information for the two soft-masks.

E. Implementation

- *Training*. Given an unlabeled audio track convolved with a Binaural Room Impulse Response (BRIR), the low-level features can be calculated as in section II-B and are used in the input layer. The speech tracks are generated by convolving a certain number of sentences from the TIMIT dataset [14] with the BRIRs captured in real echoic rooms [15], which consists of several full band audio samples recorded by a sensor placed around a half-circular grid, in

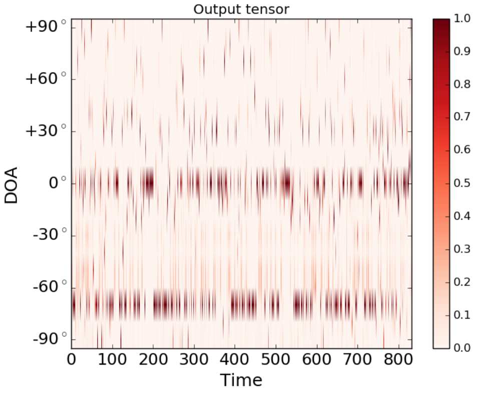


Fig. 4: Example of DNN output for test set $j = 2$. By averaging over the time frame, two DOAs can be identified as the most probable, $-70°$ and $0°$.

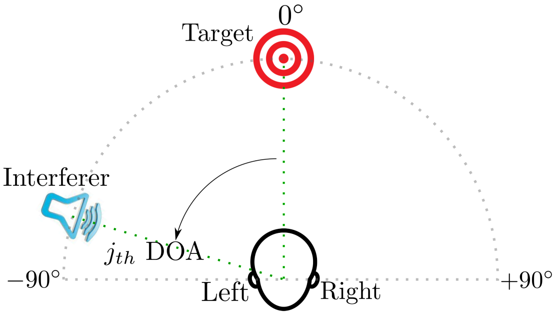


Fig. 5: The experimental setup.

variable positions ranging from $-90°$ to $+90°$, with steps of $10°$, as shown in Figure 5. White Gaussian noise with $SNR = 100\ dB$ has been added to the audio recordings in the training set.

A back-propagation algorithm is used in order to minimize the cost-function and to find the global optimized parameters for the whole deep network. The cost-function used is the cross-entropy, given by

$$\mathcal{L} = -\frac{1}{M}\left[\sum_{m=1}^{M}\sum_{j=1}^{J} 1\{y^{(m)} = j\}\log\frac{e^{W_j^T x^{(m)}}}{\sum_{l=1}^{J} e^{W_l^T x^{(m)}}}\right].$$

where W stands for the DNN weights, x is the input data and j indicates the labels.

The ground-truth for the softmax classifier is obtained from the orientation information of the unlabeled data. If the individual source in the observed signals belongs to the DOA j, $p(y_j = j|\vec{x}_{(n,m)}) = 1$ otherwise $p(y_j \neq j|\vec{x}_{(n,m)}) = 0$.

- *Testing*. Soft-masks can be generated with the set of training parameters $(\vec{W}, \vec{b})$ and used to estimate the audio sources from the mixtures.

III. Experiments

A. Experimental setup

The binaural audio recordings were simulated by convolving the target speech and the interferer with associated BRIRs, as shown in Fig. 5. The BRIR dataset was recorded around a half-circular grid, ranging from $-90°$ to $90°$ with steps of $10°$, for a total of $J = 19$ DOAs. For the training set, a total of 8 sentences, 4 for each gender and randomly selected dialects, have been concatenated and convolved with the rooms BRIRs, for a total of $26\ s \times 19\ DOAs$. The room characteristics can be seen in Table I, where room 'A' is the least reverberant and 'D' the most reverberant.

Room	Type	ITDG(ms)	DRR(dB)	$T60(s)$
A	Medium office	8.73	6.09	0.32
D	Large seminar room	21.6	6.12	0.89

TABLE I: Room acoustic properties.

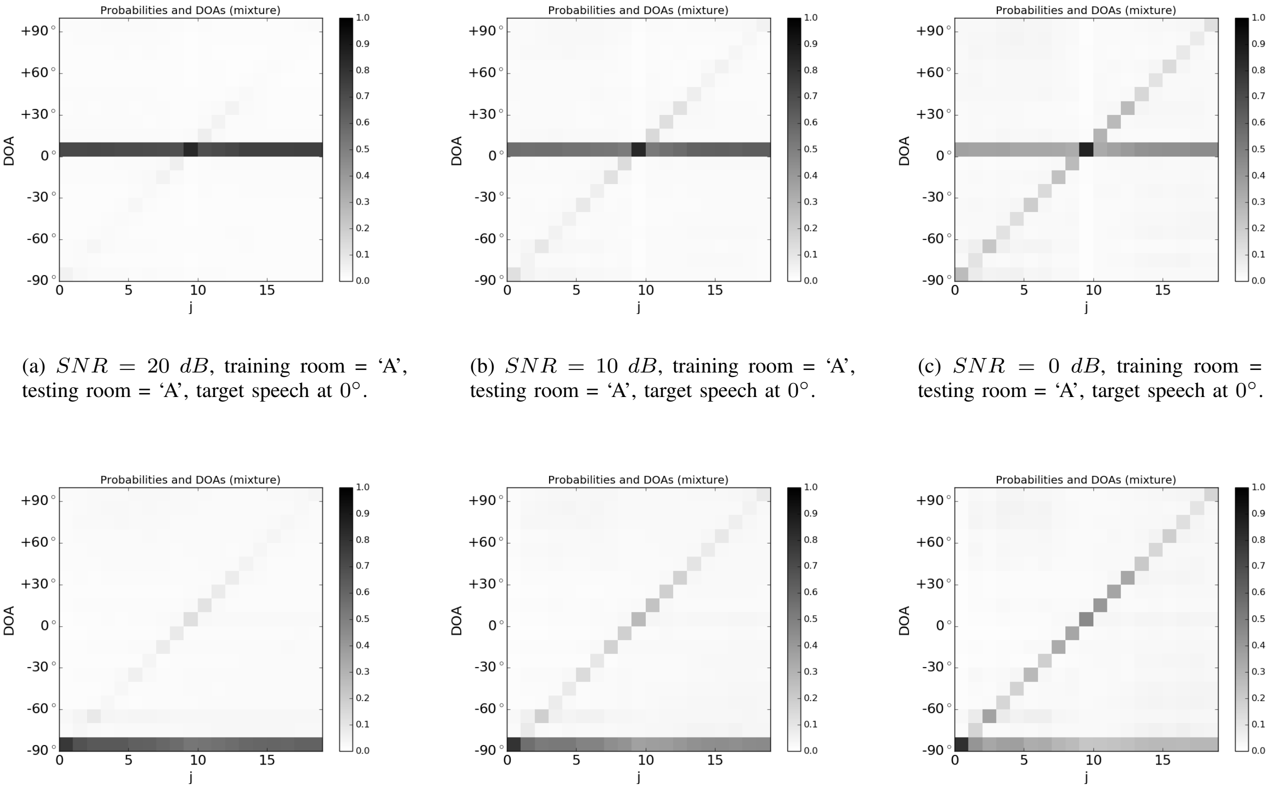
Both the target and interferer are located 1.5 m away from the dummy head and had the same height as the dummy head. In the experimental setup, the target is fixed at $0°$ or $-90°$ while the interferer is located into the different azimuthal positions. The speech-noise mixtures are built in a similar way as the training set, 10 sentences for each gender and randomly selected dialects are concatenated and convolved with the BRIRs, which correspond to $J = 19$ azimuthal positions. Different kind of noise from a database of 100 types of noise used in [16] has been used as the interferer, which is equivalent to assuming superposition of their respective sound fields. [12]

The audio files are sampled at $f_s = 16kHz$ while, regarding the STFT settings, the Hann window is set to 2048 ($128ms$) samples with 75% overlap between the neighboring windows for the STFT. The number of DNNs is $N = 128$.

The number of low-level features is 3, so there are $K \times 3 = 24$ neurons for the input layers. The DNNs need exactly $J = 19$ neurons for the output layer, each one corresponding to the number of slices in which the space is subdivided. While in the training phase the output layer is preset to the ground-truth, during the testing stage it has to be estimated given the testing inputs. Two hidden layers have been used, with $V = 1024$ units each, this has shown to achieve the best optimization. Each of the N DNNs is trained in 700 epochs, with a batch size of 400. Each DNN is trained by using the back-propagation algorithm.

B. DOA detection

Figure 6 represents the estimated DOAs as a function of the test set j, for several SNRs. In the upper row the speech target has been placed at $0°$ while in the lower row the target speech is fixed at $-90°$. Each bin represents the probability that the DOA for the target speech or the interferer noise is estimated correctly, the sum of the probabilities along each test set is normalized to 1. Figures 6 (a) and (b) are, respectively, the



(a) $SNR = 20\ dB$, training room = ‘A’, testing room = ‘A’, target speech at $0°$.

(b) $SNR = 10\ dB$, training room = ‘A’, testing room = ‘A’, target speech at $0°$.

(c) $SNR = 0\ dB$, training room = ‘A’, testing room = ‘A’, target speech at $0°$.

(d) $SNR = 20\ dB$, training room = ‘A’, testing room = ‘A’, target speech at $-90°$.

(e) $SNR = 10\ dB$, training room = ‘A’, testing room = ‘A’, target speech at $-90°$.

(f) $SNR = 0\ dB$, training room = ‘A’, testing room = ‘A’, target speech at $-90°$.

Fig. 6: Estimated DOA as a function of the test set j, interferer position at different DOAs, for the case with ILD+IPD+LPS.

cases of SNR = $20\ dB$ and $10\ dB$ and show that the probability of the target DOA is much higher than the interferer. The estimated DOA of the target speech is $0°$ (i.e. in the center), which agrees with the experimental setup. In Figure 6 (c), the $0\ dB$ case, the interferer probability becomes similar to the target probability, causing ambiguity in the target soft-mask and it is harder to separate the source as compared to the other DOA angles. The same observation can be made for Figures 6 (d), (e) and (f), where the target speech is placed at $-90°$.

All the plots in Figure 6 show how certain bins along the target DOAs are detected, this is due to a not optimal speech recognition, which the DNNs interpret as misplaced bins. By comparing the values on the grey-scale of Figure 6 (a) and Figure 7, which correspond to the $SNR = 20\ dB$ case, it can be noticed how the additional LPS feature helps generating a slightly higher probability of ~ 0.1 for the target speech DOA and, at the same time, it reduces the probability to get misplaced bins from the other directions. This would lead to a better estimation of the target DOAs and, consequently, to a better soft-mask.

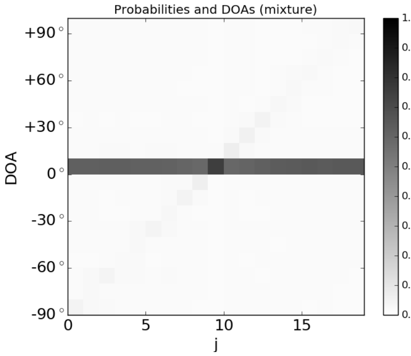
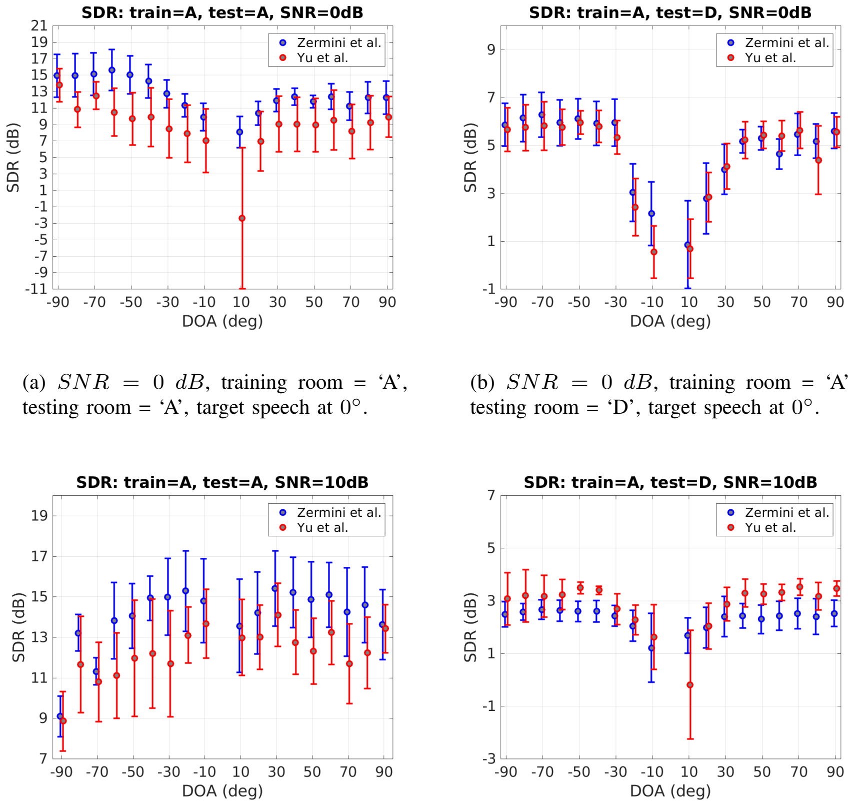


Fig. 7: Estimated DOA vs test set j, ILD+IPD, $SNR = 20\ dB$, training room = ‘A’, testing room = ‘A’, target at $0°$.

C. SDRs evaluation

Different kinds of noise at different levels of SNRs have been tested for the intereferer source. Here the results for the $0\ dB$ and $10\ dB$ SNR cases will be presented, respectively when the target is located at $0°$ and $-90°$. The test noises chosen for the presented results are looped several times in order to reach



(a) $SNR = 0\ dB$, training room = ‘A’, testing room = ‘A’, target speech at 0°.

(b) $SNR = 0\ dB$, training room = ‘A’, testing room = ‘D’, target speech at 0°.

(c) $SNR = 10\ dB$, training room = ‘A’, testing room = ‘A’, target speech at $-90°$.

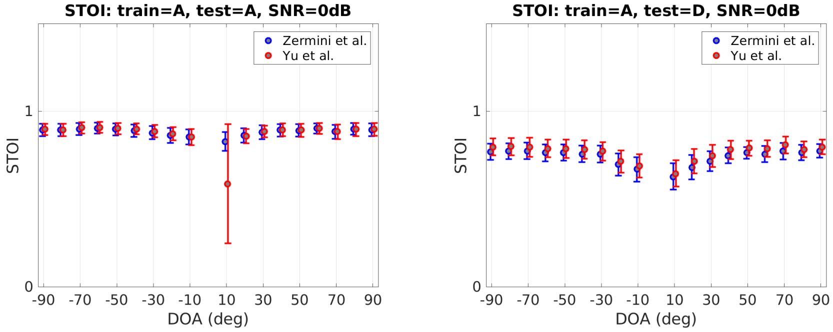
(d) $SNR = 10\ dB$, training room = ‘A’, testing room = ‘D’, target speech at $-90°$.

Fig. 8: SDRs comparisons as a function of the DOA.

the same length of the speech recording. Figure 8 shows the SDRs plot for the noise in the case in which the same room, labelled ‘A’, is used for training and the DNNs are applied for separating the speech-noise mixtures in reverberant rooms labelled ‘A’ and ‘D’, where the latter is more challenging. The points and the bars indicate, respectively, the average values and the standard deviations for each DOA. In Figures 8 (a) and (c), it can be observed that in the case in which the DNNs are trained with the proposed method, which consists in using the ILD, IPD and LPS and tested in room ‘A’, that the SDRs range from $\sim 8\ dB$ to $\sim 16\ dB$ and $\sim 9\ dB$ to $\sim 15\ dB$, respectively if the target is fixed at 0° and $-90°$. In comparison, if the DNNs are trained with the method proposed by Yu et al. [8], [9], using only the binaural low-level features ILD and IPD, the SDRs are on average $\sim 3\ dB$ lower. The training and test sets are the same in both cases. Figures 8 (b) and (d) show the case of testing using room ‘D’, which has a longer reverberation time. Using the LPS feature increases the SDRs up to $\sim 1\ dB$ for testing room ‘A’ only for certain DOAs. When the testing room is ‘D’, the proposed method has a slightly worse performance. However, in Figures 8 (a) and (d), it can be observed how at $+10°$, the SDR is negative when training with the method by Yu et al., while it is positive when the additional LPS feature is introduced. This is due to the fact that the DNNs system may sometimes interpret the target as the interferer and vice-versa, due to a poor DOA estimation and the proposed method helps to correctly classifying the two audio sources. The lower SDRs in Fig. 8 (c) and (d) compared to Fig. 8 (a) and (b) can be explained by considering that the target contribution arriving at the far-side ear is attenuated as compared to that of the near-side ear, resulting in less effective binaural features. Figures 9 (a) and (b) show the STOI (Short-Time Perceptual Intelligibility) [17] for the separated speech in room ‘A’ and ‘D’, when the target is placed at 0°. The evaluation over the two cases with and without LPS gives comparable intelligibility scores.

IV. Conclusions and Future Work

We have presented the results of a system of DNNs, trained with different low-level features and applied to a mixture of speech and noise to restore the original speech. The binaural features, the ILD and the IPD, already tested in [18] for the case of speech-speech separation, can also be applied to the case of noisy speech mixtures to achieve good separation results in terms of SDR. The LPS can be a useful additional feature when the testing room has a relatively short



(a) $SNR = 0\ dB$, training room = 'A', testing room = 'A', target speech at $0°$.

(b) $SNR = 0\ dB$, training room = 'A', testing room = 'D', target speech at $0°$.

Fig. 9: STOIs comparisons as a function of the DOA.

reverberation time while, for more reverberant rooms, the improvement is small or negligible. A possible explanation could be that the early reflections are more correlated with the direct signal, so the LPS seems to be more effective where the reflections are less, unless the reverb for speech and noise are different. To summarize the results, the method presented in this paper performs better than the baseline method, which consists in training with binaural features only [8], [9], with improvements up to 3 dB for room with short reverberation times and less significant improvements for rooms with longer reverberation times. One more advantage of this method is the fact that, compared to other works such as [10] [11], where a large amount of varied training data has been used, good separation results can be achieved by training with a small amount of training data. These works indicate that inserting a significant amount of noise information in the training data may allow the DNNs to learn how to better recognize the noise. Further work will be carried out to improve the DOAs estimation and consider training with different rooms.

V. Acknowledgements

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7-PEOPLE-2013-ITN) under grant agreement no 607290 SpaRTaN.

References

[1] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation : Independent Component Analysis and Applications*, ser. Communications engineering. Amsterdam, Boston (Mass.): Elsevier, 2010. [Online]. Available: http://opac.inria.fr/record=b1130538

[2] D. Wang and G. Brown, "Computational auditory scene analysis: Principles, algorithms, and applications." in *IEEE Transactions on Neural Networks*, 2008, p. 199.

[3] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

[4] D. D. Lee and H. S. S., "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems 13*, T. Leen, T. Dietterich, and V. Tresp, Eds. MIT Press, 2001, pp. 556–562. [Online]. Available: http://papers.nips.cc/paper/1861-algorithms-for-non-negative-matrix-factorization.pdf

[5] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[6] Y. Jiang, D. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 22, no. 12, pp. 2112–2121, Dec. 2014. [Online]. Available: http://dx.doi.org/10.1109/TASLP.2014.2361023

[7] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, Dec. 2015.

[8] Y. Yu, W. Wang, and P. Han, "Localization based stereo speech source separation using probabilistic time-frequency masking and deep neural networks," in *EURASIP Journal on Audio Speech and Music Processing*, Sept 2016, pp. 7–18.

[9] Y. Yu and W. Wang, "Unsupervised feature learning for stereo source separation," in *Proc. 10th International Conference on Mathematics in Signal Processing (IMA 2014), Birmingham, UK*, Dec 2014, pp. 15–17.

[10] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 23, no. 1, pp. 7–19, Jan. 2015. [Online]. Available: http://dx.doi.org/10.1109/TASLP.2014.2364452

[11] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, Jan 2014.

[12] A. Alinaghi, P. J. B. Jackson, Q. Liu, and W. Wang, "Joint mixing vector and binaural model based stereo source separation." in *IEEE/ACM Transactions on Audio, Speech & Language Processing*, vol. 22, no. 9, 2014, pp. 1434–1448.

[13] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization." in *IEEE Transactions on Audio, Speech & Language Processing*, 2010, pp. 382–394.

[14] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA NIST TIMIT acoustic-phonetic continous speech corpus CD-ROM," in *NASA STI/Recon Technical Report N, vol. 93*, 1993, p. 27403.

[15] C. Hummersone, "A psychoacoustic engineering approach to machine sound source separation in reverberant environments," https://github.com/IoSR-Surrey/RealRoomBRIRs/, 2011.

[16] Y. Xu, J. Du, Z. Huang, L. Dai, and C. Lee, "Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[17] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2010, pp. 4214 – 4217.

[18] A. Zermini, Y. Yu, Y. Xu, W. Wang, and M. D. Plumbley, "Deep neural network based audio source separation," in *Proc. 11th International IMA International Conference on Mathematics in Signal Processing*, 2016.