MULTICHANNEL SPEECH SEPARATION WITH RECURRENT NEURAL NETWORKS FROM HIGH-ORDER AMBISONICS RECORDINGS

Lauréline Perotin[⋆†] *Romain Serizel*[†] *Emmanuel Vincent*[†] *Alexandre Guérin*[⋆]

[⋆] Orange Labs, 4 rue du Clos Courtel, BP 91226, 35512 Cesson-Sévigné, France
[†] Université de Lorraine, CNRS, Inria, Loria, F-54000 Nancy, France

ABSTRACT

We present a source separation system for high-order ambisonics (HOA) contents. We derive a multichannel spatial filter from a mask estimated by a long short-term memory (LSTM) recurrent neural network. We combine one channel of the mixture with the outputs of basic HOA beamformers as inputs to the LSTM, assuming that we know the directions of arrival of the directional sources. In our experiments, the speech of interest can be corrupted either by diffuse noise or by an equally loud competing speaker. We show that adding as input the output of the beamformer steered toward the competing speech in addition to that of the beamformer steered toward the target speech brings significant improvements in terms of word error rate.

Index Terms— Speech separation, high-order ambisonics (HOA), multichannel filtering, LSTM

1. INTRODUCTION

Audio source separation is the process of extracting one or several sources from an audio mixture. It is of particular interest as a pre-processing step for automatic speech recognition (ASR) in adverse contexts such as distant voice command or automatic meeting transcription. Many methods rely on multichannel filtering, which is known to introduce less speech distortion than single-channel filtering, thereby facilitating ASR [1]. Approaches include multichannel Wiener filtering (MWF) and its variants such as the speech distortion weigthed MWF [2,3] and corresponding implementations based on the eigenvalue decomposition (EVD) or the generalized EVD (GEVD) [4–6]. The computation of these filters generally relies on the assumption that the acoustic sources are uncorrelated and the availability of a speech presence probability estimator.

The application of deep neural networks (DNNs) to source separation has allowed for drastic improvement of ASR accuracy in real-world conditions [7]. DNNs were originally applied to single-channel inputs to derive a single-channel filter, a.k.a. a mask [8–10]. In the multichannel case, several approaches have been proposed to pass spatial information directly to a DNN, for instance using phase difference features between non-coincident microphones [11] or coherence features [12]. However, in these two studies, the mask estimated by the DNN is still applied as a single-channel filter only. Recent approaches that derive DNN-based multichannel filters have proven very promising [7, 13]. These include various beamformers derived from the speech and noise covariance matrices computed from the output mask [7, 14] or an MWF derived by expectation-maximization [13]. Yet these approaches target a single speaker in a noisy environment while real-world scenarios often involve several speakers. Deep clustering [15] was proposed as way to solve the multi-speaker problem but it is currently limited to single-channel processing and involves a significant computational cost.

We are targeting voice command applications in enhanced reality environments for improved user experience. Therefore we consider inputs in the high order ambisonics (HOA) format [16], which represents the spherical harmonics decomposition of the sound field at a given point in space. This format can be obtained from a variety of spherical arrays and is hence independent of the recording device. It is becoming increasingly popular in industrial applications of enhanced or virtual reality such as Youtube360 [1] or Facebook360 [2] owing to the fact that it is isotropic and provides a representation of spatial audio scenes that is very convenient to manipulate or respatialize. In order to do such manipulations, one must have access to the source signals. Very few studies have focused on HOA source separation so far [17,18] and none of them exploits the recent advances in the domain thanks to DNN.

In this paper, we propose a recurrent neural network based multichannel source separation algorithm and show its application to HOA inputs. In addition to the classic situation of a single speaker recorded in diffuse noise, we consider the challenging case of concurrent speakers with the same intensity. As some information is necessary to identify the target, we assume that the source directions of arrival (DoAs) are known. Indeed, algorithms providing a reliable estimation of the DoAs of HOA signals already exist [19].

Section 2 defines the notations and the signal model, and

1. https://support.google.com/youtube/answer/6395969
2. https://facebookincubator.github.io/facebook-360-spatial-workstation/KB/CreatingVideosSpatialAudioFacebook360.html

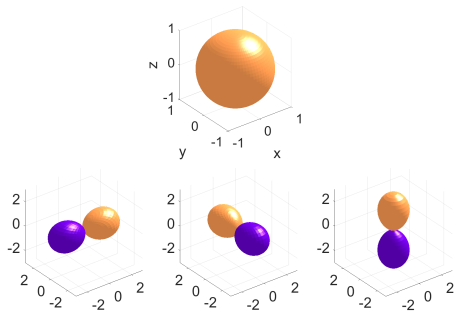

Fig. 1. Power of the first spherical harmonics for order 0 (top) and 1 (bottom). Light parts correspond to the area where the spherical harmonic is positive, while darker parts correspond to the negative area.

reminds the fundamentals about the HOA format and MWF. In Section 3, we present our DNN-based multichannel source separation approach for HOA. The experimental setup is presented in Section 4 and the results are discussed in Section 5. Finally, we conclude in Section 6.

2. PROBLEM STATEMENT

2.1. Signal model

We consider a mixture of target speech and noise as recorded by N microphones. In the short-time Fourier transform (STFT) domain, the $N \times 1$ vector $\mathbf{x}(t,f)$ of mixture STFT coefficients can be written as

$$\mathbf{x}(t,f) = \mathbf{s}(t,f) + \mathbf{n}(t,f) \tag{1}$$

where t and f are the time frame and bin indexes, $\mathbf{s}(t,f)$ is the $N \times 1$ spatial image of the target speech recorded at the microphones, and the $N \times 1$ noise vector $\mathbf{n}(t,f)$ can contain J competing directional speakers and diffuse noise.

2.2. High-order ambisonics

The HOA format decomposes the sound field at a particular point of space on the basis of spherical harmonics functions. The HOA channels are the coefficients of the decomposition on each function of the basis. The spherical harmonics functions are grouped by order. Order 0 involves one simple omnidirectional function. The number and the complexity of the functions increase with the order (see Fig. 1). The decomposition is only exact when using the full infinite spherical harmonics basis.

In the following, we will use only the $N = 4$ channels of the first-order decomposition which can be considered as virtual microphones. It will prove sufficient for speech enhancement (see Section 5). Besides, this decomposition can be obtained with a limited number of microphones and requires limited computations, making it suitable for implementation in embedded systems. The four channels we use, traditionnally named W, X, Y and Z, are ideally what would record an omnidirectionnal microphone (order 0) and three perfect bidirectional microphones pointing towards the axes X, Y and Z (order 1), all four being coincident (see Fig. 1). For a point source $p(t,f)$ assimilated to a plane-wave coming from azimuth θ and elevation ϕ, the decomposition of the sound field on these four channels is given by the frequency-independent HOA steering vector $\mathbf{d}_{\theta,\phi}$:

$$\begin{bmatrix} W(t,f) \\ X(t,f) \\ Y(t,f) \\ Z(t,f) \end{bmatrix} = \mathbf{d}_{\theta,\phi}\, p(t,f) = \begin{bmatrix} 1 \\ \sqrt{3}\cos\theta\cos\phi \\ \sqrt{3}\sin\theta\cos\phi \\ \sqrt{3}\sin\phi \end{bmatrix} p(t,f). \tag{2}$$

Capitalizing on this representation, a simple anechoic beamformer can be derived for the $J+1 < N$ directional sources by inverting the matrix of steering vectors, to point toward the DoA of source i and cancel sounds coming from the J other DoAs [19] :

$$\mathbf{b}_i = [\mathbf{d}_{\theta_0,\phi_0}\, \mathbf{d}_{\theta_1,\phi_1} \ldots \mathbf{d}_{\theta_J,\phi_J}]^{\dagger}\mathbf{u_i}. \tag{3}$$

† designates the pseudo-inverse and $\mathbf{u}_i$ is the vector selecting the i-th row of a matrix (only zeros except a one in the i-th position). According to the mixture model (1), b_0 points toward the target and $b_{i\geq 1}$ toward the directional interferences. This beamformer can be applied in the case with only one directional source and is then similar to a simple delay-and-sum beamformer.

2.3. Mutichannel Wiener filters

In reverberant conditions, the above beamformer achieves limited enhancement performance. Instead, an MWF $\mathbf{w}(f)$ is applied to the mixture $\mathbf{x}(t,f)$ to obtain

$$y(t,f) = \mathbf{w}(f)^H \mathbf{x}(t,f). \tag{4}$$

Wang et al. [14] studied several variants of the MWF for ASR purposes. The most promising one is derived from a rank-1 approximation of the MWF based on the GEVD. Let $\mathbf{\Phi_{ss}}(f)$ and $\mathbf{\Phi_{nn}}(f)$ be the $N \times N$ covariance matrices of speech and noise and $\mathbf{\Phi_{ss-r1}}(f)$ the rank-1 approximation of $\mathbf{\Phi_{ss}}(f)$:

$$\mathbf{\Phi_{ss-r1}}(f) = \sigma(f)\mathbf{a}(f)\mathbf{a}(f)^H \tag{5}$$

with $\sigma(f)$ the largest eigenvalue of $\mathbf{\Phi_{nn}}(f)^{-1}\mathbf{\Phi_{ss}}(f)$ and $\mathbf{a}(f)$ the associated eigenvector. The GEVD-MWF [6] is then :

$$\mathbf{w}_{\mathrm{GEVD}}(f) = [\mathbf{\Phi_{ss-r1}}(f) + \mathbf{\Phi_{nn}}(f)]^{-1}\mathbf{\Phi_{ss-r1}}(f)\mathbf{u}_1. \tag{6}$$

In practice, the covariance matrix of the target speech is estimated by averaging an estimate $\tilde{\mathbf{s}}(t,f)$ of the signal $\mathbf{s}(t,f)$

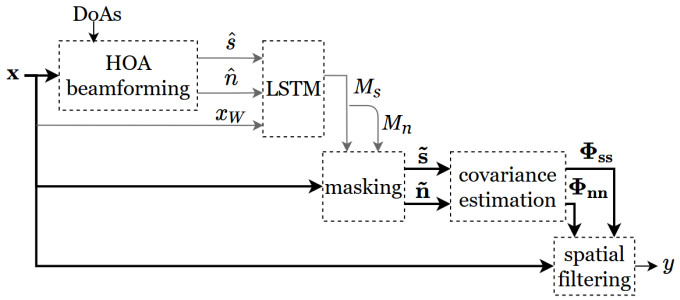

Fig. 2. Proposed separation system.

over all time frames of the utterance :

$$\mathbf{\Phi_{ss}}(f) = \frac{1}{T}\sum_{t=0}^{T-1} \tilde{\mathbf{s}}(t,f)\tilde{\mathbf{s}}^H(t,f) \quad (7)$$

$\tilde{\mathbf{s}}(t,f)$ can be obtained by applying a time-frequency mask $M_s(t,f)$ estimated via a DNN [7, 14] :

$$\tilde{\mathbf{s}}(t,f) = M_s(t,f)\mathbf{x}(t,f). \quad (8)$$

The noise covariance matrix $\mathbf{\Phi_{nn}}(f)$ is obtained similarly.

3. STRUCTURE OF THE SOLUTION

The proposed solution is described in Fig. 2. We use a long-short term memory (LSTM) recurrent neural network to estimate time-frequency masks. In the training stage, the target speech mask M_s is computed from the magnitude spectra $s_{\mathrm{W}}(t,f)$ and $n_{\mathrm{W}}(t,f)$ of the target speech and of the noise observed in the omnidirectional channel :

$$M_s(t,f) = \frac{s_{\mathrm{W}}(t,f)^2}{s_{\mathrm{W}}(t,f)^2 + n_{\mathrm{W}}(t,f)^2}. \quad (9)$$

The noise mask is then deduced as $M_n(t,f) = 1 - M_s(t,f)$. The masks are used to compute the covariance matrices $\mathbf{\Phi_{ss}}(f)$ and $\mathbf{\Phi_{nn}}(f)$. The GEVD-MWF is finally obtained from these covariance matrices according to (6).

The choice of the inputs to the LSTM can have a great impact on the system's performance (see Section 5). In the case of competing speakers, it is necessary that the network be given some additional information to identify the target. In previous works, the input of the network was either the magnitude spectrum of the mixture [7] or of the mixture processed with a simple delay-and-sum beamformer [13]. We propose to combine the magnitude spectra of the mixture observed at the omnidirectional channel W, $x_W(t,f)$, and of the ouput of the HOA beamformer pointing toward the target, $\hat{s}(t,f)$:

$$\hat{s}(t,f) = |\mathbf{b_0}^H\mathbf{x}(t,f)|. \quad (10)$$

Additionally, we combine these inputs with the magnitude spectra of the outputs of the beamformers pointing toward each interference with known DoA, $\hat{n}_i(t,f), i \in \{1,..,J\}$:

$$\hat{n}_i(t,f) = |\mathbf{b}_i^H\mathbf{x}(t,f)|. \quad (11)$$

Such explicit information about the noise was already used in traditional filtering with the generalized sidelobe canceller [20] but, to the best of our knowledge, it had not been used in DNN-based multichannel speech separation so far.

4. EXPERIMENTAL SETUP

4.1. Data

We conducted our experiments in two situations : with a single speaker and with two speakers.

For training, the target speech was picked among 1801 utterances of 10 s duration from the French newspaper read speech corpus Bref [21]. It was convolved with HOA spatial room impulse responses (SRIR) measured in a room with TR60 = 270 ms, generated by 16 loudspeakers regularly positioned in space, at 2.30 m distance from the central microphone. For each utterance, one SRIR was randomly picked among those 16. The validation set is made of 684 utterances from Bref convolved with SRIRs recorded in a different but comparable room, using different speakers. The test speech utterances come from the Ester dataset [22], made of real French television or radio contents. We extracted 20 sentences of 1 minute containing only speech (without music or jingles), for a total of 4043 words. The SRIRs were measured in a room with TR60 = 350 ms and a microphone in the center at 1.65 m distance from the loudspeakers. We added babble noise to all the speech signals, randomly picked from Freesound [3] in different subsets for training, validation and test. In order to simulate diffuse noise, we convolved this noise with the mean of the diffuse part of two randomly chosen SRIRs.

In the single-speaker case, the signal-to-noise ratio (SNR) is 0 dB. In the two-speaker case, the SRIRs for the target and the competing speakers are picked in the median plane with various azimuth differences (25°, 45° or 90°). The signal-to-interference ratio (SIR) between the target and the competing speech is 0 dB. The babble noise is added with 20 dB SNR.

4.2. Setup

All the data are sampled at 16 kHz. We compute the STFT with a sinusoidal window of 1024 samples and 50% overlap. We use a network made of one LSTM layer with 512 hidden units and the tanh activation function, plus one output feed-forward layer with 513 units and the sigmoid activation so as to predict mask coefficients between 0 and 1. The LSTM takes as input sequences of 25 frames with 12 overlapping frames between two sequences.

The network was trained against the mean-square error cost function with the Nadam optimizer [23]. The learning

3. http://freesound.org

			1 spk	2 spk, angle diff.		
				25°	45°	90°
Clean speech			7.4	7.4	7.4	7.4
Mixture			68.5	91.7	88.9	85.4
Beamformer (3)			24.3	76.0	45.9	20.6
Ideal mask (9)			18.3	16.3	15.0	16.3
Filter from ideal mask			13.1	23.0	16.5	11.1
Network inputs	x_W	mask	68.6	91.8	84.5	85.7
		filter	25.0	91.6	87.1	86.6
	$\hat{s}$	mask	61.2	90.8	84.8	78.3
		filter	19.6	67.2	27.1	12.9
	x_W, $\hat{s}$	mask	55.9	86.4	61.6	45.0
		filter	**17.1**	80.9	21.0	**10.5**
	x_W, $\hat{s}$, $\hat{n}_1$	mask	n/s	60.9	43.9	37.2
		filter		**22.3**	**14.5**	**11.0**

Table 1. WER (%) achieved on the reference signals (top) and on the outputs of our method depending on the network inputs (bottom). "mask" corresponds to the first channel of $\tilde{\mathbf{s}}$ in (8) and "filter" to y, the output of the GEVD-MWF in (4). The best result in each situation is shown in bold. When the confidence intervals of two results overlap, both are shown in bold.

rate is initially set to 10^{-3}. We use a 10^{-4} L2 weight regularization as well as 50% dropout, both for recurrent and non-recurrent weights. The number of iterations, controlled by early stopping, is limited to 10.

4.3. ASR evaluation

Performance is measured by the word error rate (WER) obtained by Cobalt Speech Recognition, developed by Orange Labs for French ASR. It is a Kaldi-based speech-to-text decoder using a time-delay neural network based acoustic model [24] trained on more than 2000 h of clean and noisy speech, a 1.7-million-word lexicon, and a 5-gram language model trained on 3 billion words. Given the size of our test corpus, the best results achieved by our method are given with a 95% confidence interval of $\pm 1.0\%$.

Right before recognition, all signals are dereverberated using NTT's weighted prediction error based system [25]. We used 50 filter coefficients and a prediction delay of 3, as this setting ensures a soft dereverberation with limited distortion.

5. RESULTS

All results are summarized in Table 1. To our knowledge, there are no previous source separation systems for HOA contents aiming at ASR. Therefore, the baseline we use for comparison is the output of the beamformer in (3). The lower bound is the clean target speech signal (with no reverberation) : it indicates the WER due to the ASR system itself. Improving this lower bound is outside the scope of this article. We also compare the results to those achieved with the GEVD-MWF computed from the ideal mask (9). This tells us how much improvement might still be achievable by improving the mask returned by the neural network.

To compare our results with the single-channel state of the art, we first give our network one input only. Given the omnidirectional channel x_W, the system is able to use the different spectral characteristics of speech and babble noise to improve the WER compared to the mixture for a single speaker, but no more than the baseline. However, it is not enough for the network to guess which is the target in the two-speaker scenarios. On the other hand, when $\hat{s}$ (10) is given, the mask returned by the network allows the GEVD-MWF to improve the WER compared to the baseline in all cases, for instance by a relative 37% when the speakers are 90° apart and 41% when they are 45° apart.

When those inputs are combined, the results further improve in the single-speaker case, reaching a relative WER reduction of 30% compared to the baseline, as well as for the two speakers with 90° or 45° angle difference. However, when the sources are too close for the beamformer to discriminate (25°), the mask returned by the network is of no help for the multichannel filter, which performs as badly as the simple beamformer and barely improves compared to the mixture.

This can be overcome by feeding the network with $\hat{n}_1$ (11). When the speakers are 90° apart, this does not significantly improve the already good performance of the system. Nevertheless, in the case of two speakers that are 25° apart, adding this information brings a huge improvement : the WER becomes 71% relative better than the baseline. [4]

6. CONCLUSION

In this paper we presented a new speech separation system and showed its efficiency on HOA contents. It uses LSTM-based mask estimation to compute a GEVD multichannel Wiener filter. Given knowledge of the DoAs, we showed that in the most difficult case (equally loud competing speakers coming from close directions) inputting the network with the information given by simple beamformers pointing toward the target and the interference, respectively, can radically improve the WER performance. In future works, we intend to check the robustness of the system against small errors in the estimated DoAs.

4. Audio examples are available on : `https://members.loria.fr/LPerotin/demos/icassp2018`.

7. REFERENCES

[1] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, et al., "The NTT CHiME-3 system : Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. of ASRU*, 2015, pp. 436–443.

[2] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction," *Speech Communication*, vol. 49, no. 7, pp. 636–656, 2007.

[3] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 18, no. 2, pp. 260–276, 2010.

[4] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. on Sig. Proc.*, vol. 50, no. 9, pp. 2230–2244, 2002.

[5] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 15, no. 5, pp. 1529–1539, 2007.

[6] R. Serizel, M. Moonen, B. Van Dijk, and J. Wouters, "Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 22, no. 4, pp. 785–799, 2014.

[7] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. of ICASSP*, 2016, pp. 196–200.

[8] F. Weninger, F. Eyben, and B. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *Proc. of ICASSP*, 2014, pp. 3709–3713.

[9] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. of ICASSP*, 2015, pp. 708–712.

[10] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. of GlobalSIP*, 2014, pp. 577–581.

[11] P. Pertila and J. Nikunen, "Distant speech separation using predicted time-frequency masks from spatial features," *Speech Communication*, vol. 68, pp. 97–106, 2015.

[12] A. Schwarz, C. Huemmer, R. Maas, and W. Kellermann, "Spatial diffuseness features for DNN-based speech recognition in noisy and reverberant environments," in *Proc. of ICASSP*, 2015, pp. 4380–4384.

[13] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 24, no. 9, pp. 1652–1664, 2016.

[14] Z. Wang, E. Vincent, R. Serizel, and Y. Yan, "Rank-1 constrained multichannel Wiener filter for speech recognition in noisy environments," *arXiv :1707.00201 [cs]*, 2017, arXiv : 1707.00201.

[15] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering : Discriminative embeddings for segmentation and separation," in *Proc. of ICASSP*, 2016, pp. 31–35.

[16] J. Daniel, *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia*, Thèse de doctorat, Univ. Paris VI, France, 2000.

[17] N. Epain and C. Jin, "Independent component analysis using spherical microphones arrays," *Acta Acustica united with Acustica*, vol. 1, no. 98, pp. 91–102, 2012.

[18] P. K. T. Wu, N. Epain, and C. Jin, "A super-resolution beamforming algorithm for spherical microphone arrays using a compressed sensing approach," in *Proc. of ICASSP*, 2013, pp. 649–653.

[19] M. Baqué, *Analyse de scène sonore multi-capteurs*, Ph.D. thesis, Univ. du Maine, 2017.

[20] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas. Propag.*, vol. 30, no. 1, pp. 27–34, 1982.

[21] L. F. Lamel, J.-L. Gauvain, and M. Eskénazi, "BREF, a large vocabulary spoken corpus for French," in *Proc. of Eurospeech*, 1991, pp. 505–508.

[22] G. Gravier, J.-F. Bonastre, E. Geoffrois, S. Galliano, K. Mc Tait, and K. Choukri, "The ESTER evaluation campaign for the rich transcription of French broadcast news," in *Proc. of LREC*, 2004.

[23] T. Dozat, "Incorporating Nesterov momentum into Adam," Tech. Rep., Univ. of Stanford, 2015.

[24] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, et al., "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proc. Interspeech*, 2016, pp. 2751–2755.

[25] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 18, no. 7, pp. 1717–1731, 2010.