GEOMETRIC INFORMATION BASED MONAURAL SPEECH SEPARATION USING DEEP NEURAL NETWORK

Yang Xian, Yang Sun, Jonathon A. Chambers, Syed Mohsen Naqvi

Intelligent Sensing and Communications Research Group, Newcastle University, UK
Email:{y.xian2, y.sun29, jonathon.chambers, mohsen.naqvi}@newcastle.ac.uk

ABSTRACT

The performance of deep neural network (DNN) based monaural speech separation methods is limited in reverberant and noisy room environments. In this paper, we propose a new DNN training target which incorporates geometric information describing the target speaker and microphone to improve the performance in reverberant and noisy room environments. The experiments are based on the IEEE corpus and the NOISEX database and real impulse responses (RIRs). The objective evaluations, short-time objective intelligibility (STOI) and perceptual evaluation of speech quality (PESQ) confirm the efficiency of the proposed direct path ratio mask (DRM).

Index Terms— deep neural network, speech separation, geometric information, noisy reverberant speech mixture, direct path ratio mask

1. INTRODUCTION

Speech separation has various applications such as hearing aids and automatic speech recognition (ASR) [1]. Nevertheless, the performance of state-of-the-art of speech separation methods is limited in real reverberant and noisy room environments [2].

In the last decade, the statistical signal processing and computational auditory scene analysis (CASA) based methods are applied to solve the speech separation problem [3–7]. Nowadays, the DNN based methods are common. Jiang et al. converted the speech separation to a binary classification problem and the DNN is applied to form the ideal binary mask (IBM) [8]. The adaptive discriminative criterion is applied to the DNN, which provides a better separation performance [9]. Narayanan et al. proposed the ideal ratio mask (IRM) that is a robust training target for the DNN [10]. Wang et al. make a comparison between the different training targets and the results show the masking based targets outperform the spectral envelope based targets [2]. The IRM is more accurate than the IBM, particularly for the speech denoising problem [10]. Although, the IRM has many merits, new training targets that can better reflect the clean speech and noise are still needed.

Recently, enormous efforts have been dedicated to dereverberation and denoising. The reverberant speech essentially consists of two parts: direct path speech and reverberations. In this paper, we use geometric information to describe the target speaker and microphone to calculate the direct path impulse response, which is used to estimate the direct path speech. By using the direct path speech, we propose the DRM which improves performance in noisy and reverberant room environments.

The rest of the paper is organized as follows. In Section 2, we describe the noisy reverberant and direct path impulse response model and new DNN training target. In Section 3, the experimental setup and results are shown. In Section 4, conclusions are drawn, and future work is suggested.

2. ALGORITHM DESCRIPTION

2.1. Mixture Model and Direct Path Impulse Response

The reverberant speech mixture can be modelled as the convolution result of the speech source and impulse response as:

$$y(t) = s(t) * h(t) \tag{1}$$

where '$*$' represents the convolution operator, $y(t)$ denotes the reverberant speech, $s(t)$ represents the speech source and $h(t)$ is the impulse response. The impulse response can be divided into the direct path and reflections as:

$$h(t) = h_D(t) + h_R(t) \tag{2}$$

where $h_D(t)$ is the impulse response of the direct path and $h_R(t)$ denotes the impulse response of reflections.

The geometric information provides the distance and bearing between the speech source and the microphone, which helps to estimate direct path impulse response. The direct path impulse response, as shown in Fig. 1, is calculated as:

$$h_D(t) = \beta\delta(t-\tau) = \frac{\kappa}{d^2} cos\left(\frac{\theta}{r}\right)\delta\left(t - \frac{f_s}{C}d\right) \tag{3}$$

where β denotes the attenuation rate, δ represents the unit impulse, κ represents the attenuation per unit length in air, and d is the distance between the speech source and microphone. The parameter θ represents the angle between the

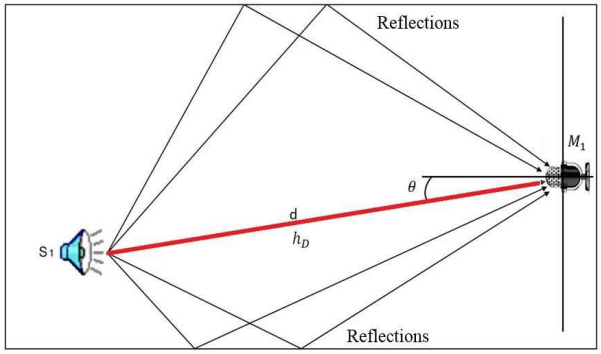

Fig. 1: Monaural speech separation setup within a reverberant room environment, the distance and angle between the target speaker and sensor are shown.

speech source and microphone, and r is the directionality coefficient. Besides, τ is the propagation time, f_s is the sample frequency, and C denotes the sound velocity in air.

Based on the distributive property of convolution, the reverberant speech mixture can be represented as [11]:

$$\begin{aligned} y(t) &= s(t) * h_D(t) + s(t) * h_R(t) \\ &= s_D(t) + s_R(t) \end{aligned} \tag{4}$$

where $s_D(t)$ is the direct path speech and $s_R(t)$ includes only reverberations. To simulate the real room environment, the reverberant speech mixture with noises is provided as:

$$y(t) = s_D(t) + s_R(t) + \alpha n(t) \tag{5}$$

where $n(t)$ denotes the noise at time t, and α is used to control the SNR level between speech and noise.

2.2. Training Targets

By using (3) and (5), the DRM can be calculated as:

$$DRM(t,f) = \left(\frac{S_D^2(t,f)}{S_D^2(t,f) + N^2(t,f)} \right)^{\eta} \tag{6}$$

where $S_D^2(t,f)$ denotes the energy of the direct path speech at time t and frequency frame f, and $N^2(t,f)$ is the energy of noise. And η is the tunable parameter to scale the mask. The proposed DRM is used as a training target, which requires less accuracy in the separation of noisy reverberant speech mixture, because the DRM mitigates reflections and noise. The direct path impulse response based speech is estimated as:

$$\hat{S}_D(t,f) = Y(t,f) DRM(t,f) \tag{7}$$

2.3. Speech Reconstruction

Since the DRM can only separate the direct path signal from the noisy reverberant mixture, the speech reconstruction module is used to separate the desired speech source. At the testing stage of the speech reconstruction module, there are two inputs: (1) the estimated direct path speech $\hat{s}_D(t)$ based on the DRM, (2) direct path impulse response h_D based on geometric information. The final time domain separated speech source is calculated as:

$$\hat{s}(t) = IFFT \left[\left(\hat{S}_D(t,f) \right) \left(H_D(t,f) \right)^{-1} \right] \tag{8}$$

where the IFFT represents the inverse fast Fourier transform operation.

2.4. System Architecture

The system architecture is shown in Fig. 2. The geometric information of the target speaker and microphone for monaural speech separation can be obtained from our multiple human tracking systems [12, 13], which are successfully used in multimodal binaural and overdetermined speech separation [14, 15]. At the training stage, the geometric information is applied to generate the proposed DRM and at the testing stage, the trained DNN with geometric information is used to estimate the final desired speech signal.

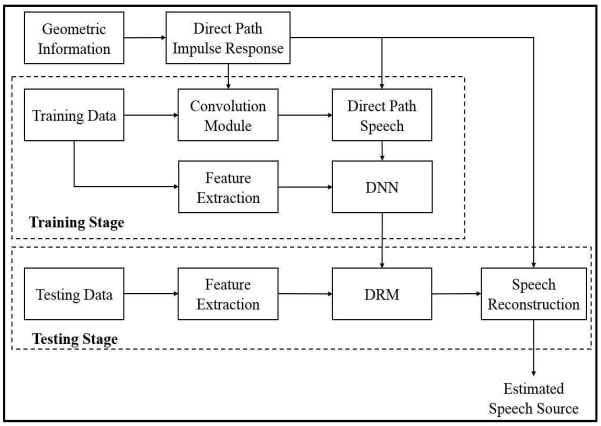

Fig. 2: The block digram of the propose reverberant and noisy speech separation system.

3. EXPERIMENTAL EVALUATIONS

3.1. Datasets

The speech signals are selected from the IEEE corpus which contains 720 utterances [16]. 500 utterances are used to genenrate the training data samples, 100 utterances are applied as development data and 120 utterances are exploited to generate the testing data. factory noise and babble noise are used as background noise, which are selected from the NOISEX datasbase, and both of them are non-stationary [17]. The direct path impulse responses are obtained by using the geometric information, which is assumed to be avaiable and can be estimated from our previous multimodal human tracking systems [13, 14]. The simulated and real room impulse

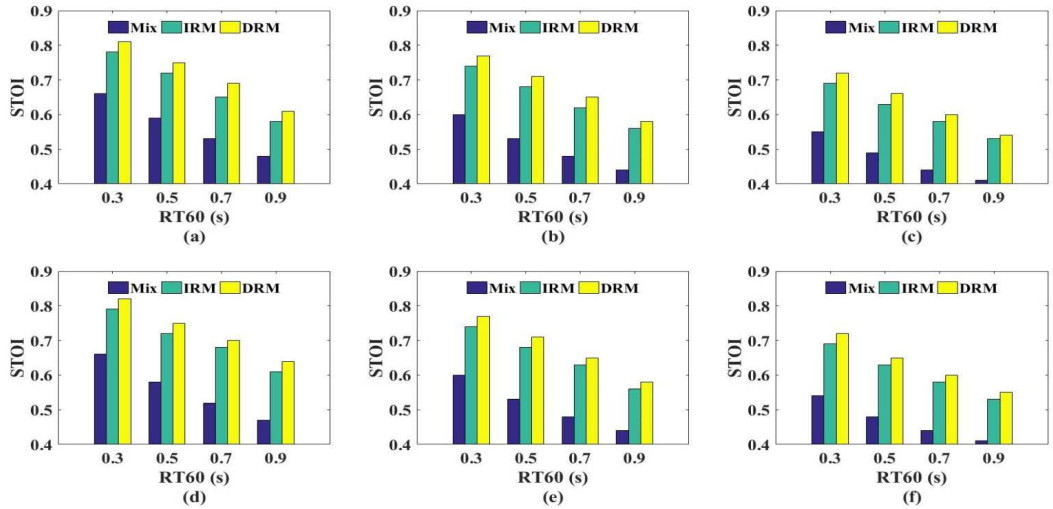

Fig. 3: Averaged STOI scores of 120 experiments for unprocessed reverberant signals, the IRM [2] and the proposed DRM systems with simulated impulse responses, subfigure:(a) 3 dB factory noise, (b) 0 dB factory noise, (c) -3 dB factory noise, (d) 3 dB babble noise, (e) 0 dB babble noise and (f) -3 dB babble noise.

responses (RIRs) are used to generate the noisy reverberant speech mixtures. The simulated RIRs are generated by the image method [18]. The room dimensions are 9 m $\times$ 5 m $\times$ 3 m, and the target source and microphone are located at 5.5 m $\times$ 2.5 m $\times$ 1.5 m and 4.5 m $\times$ 2.5 m $\times$ 1.5 m, respectively. The RT60 is increased from 0.3 s to 0.9 s with the stepsize of 0.2 s. The database recorded by Surrey University is used for the real RIRs [19], and the RT60s are 0.32 s, 0.47 s and 0.68 s. The SNR levels are set to 3 dB, 0 dB and -3 dB as in [11]. In summary for detailed evaluation of our proposed method, we have 21000 training samples, and 4200 testing samples.

The separation performance is evaluated quantitatively by two measures, they are STOI and PESQ [20, 21]. The STOI ranges from 0 to 1, where 0 means the worst intelligibility and 1 means the best intelligibility, and it has high correlation with human speech intelligibility scores [11]. The PESQ ranges from -0.5 to 4.5, where -0.5 represents the lowest perceptual evaluation of speech quality and 4.5 represents the highest quality.

3.2. DNN Settings and Speech Features

The DNN includes four hidden layers, and every hidden layer has 1024 units. The rectified linear unit (ReLU) function is used as the activation function of each unit at hidden layers and the activation function of the output unit is the sigmoid. The maximum number of epochs is 50. The dropout is applied to solve the over-fitting problem, and the rate of dropout is 0.2 [2]. The parameters of the DNN are initialized by random initialization, then they are optimized at every epoch by using adaptive subgradient descent algorithm that has 0.005 learning rate. After 50 epochs, the epoch with minimum cost function value is selected to perform the speech separation task, which is measured by the mean squared error (MSE) cost function.

SNR Level		3 dB		0 dB		-3 dB	
RT60(s)	Targets	factory	babble	factory	babble	factory	babble
0.3	Unprocessed	0.92	1.06	0.65	0.87	0.48	0.52
	IRM	2.40	2.45	1.95	2.25	1.72	2.03
	DRM	**2.49**	**2.50**	**2.05**	**2.35**	**1.83**	**2.19**
0.5	Unprocessed	0.64	0.83	0.51	0.68	0.45	0.55
	IRM	1.89	2.18	1.69	2.00	1.48	1.83
	DRM	**2.05**	**2.25**	**1.79**	**2.12**	**1.60**	**1.95**
0.7	Unprocessed	0.50	0.64	0.47	0.55	0.44	0.52
	IRM	1.74	1.92	1.55	1.74	1.31	1.62
	DRM	**1.85**	**2.11**	**1.61**	**1.94**	**1.44**	**1.78**
0.9	Unprocessed	0.40	0.60	0.35	0.47	0.31	0.41
	IRM	1.51	1.75	1.32	1.61	1.23	1.46
	DRM	**1.59**	**1.90**	**1.43**	**1.74**	**1.34**	**1.60**

Table 1: Averaged PESQ scores of 120 experiments for the IRM and the proposed DRM [2] systems at 3 dB, 0 dB and -3 dB SNR levels. The noisy reverberant speech mixtures are obtained by using the IEEE corpus and the factory and the babble background noise under simulated impulse responses. The bold numbers represent the best performance.

A complementary set of features is applied [11]. These features are mel-frequency cepstral coefficient (MFCC), spectral transform and perceptual linear prediction (RASTA-PLP) and amplitude modulation spectrum (AMS), and they are spectrum based features [2, 22]. Also, the deltas of RASTA-PLP, AMS and MFCC are appended to the features. The features are normalized to zero mean and unit variance.

3.3. Evaluations with Synthetic RIRs

The IRM is used as the benchmark. Table 1 and Fig. 3 show the PESQ and the STOI values of unprocessed and processed signals with different background noise and RT60s. In Table 1, the bold numbers represent the best separation performance.

In terms of PESQ, both the IRM and the proposed DRM

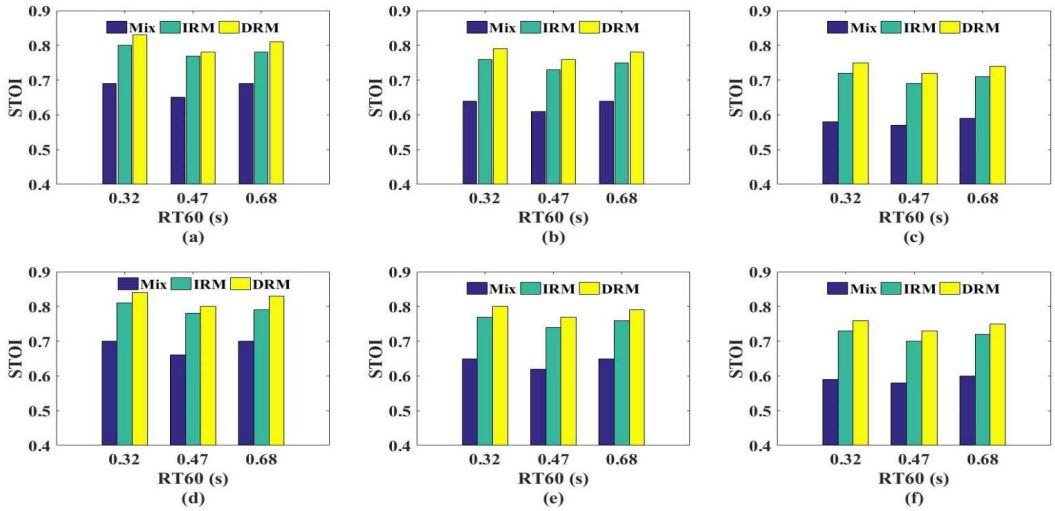

Fig. 4: Averaged STOI scores of 120 experiments for unprocessed reverberant signals, the IRM [2] and the proposed DRM systems with real impulse responses, subfigure: (a) 3 dB factory noise, (b) 0 dB factory noise, (c) -3 dB factory noise, (d) 3 dB babble noise, (e) 0 dB babble noise and (f) -3 dB babble noise.

provide considerable improvement over the unprocessed noisy reverberant signal. The proposed DRM outperforms the IRM at all RT60s. And the best PESQ performance is obtained by the DRM at the lowest RT60 (0.3 s). For example, at -3 dB SNR level with factory noise, the proposed method obtains the PESQ-improvements over the IRM as 0.16, 0.12, 0.16, 0.14 at different RT60s (0.3 s, 0.5 s, 0.7 s, 0.9 s), respectively. Because the higher RT60 increases the complexity in noisy reverberant speech mixture, the PESQ-improvement with the lower RT60 (0.3 s) is better than the higher RT60 (0.9 s). Since the noise has less effect in higher SNR levels speech mixtures, the speech separation performance will be better.

In terms of STOI scores, it is similar with the trend of PESQ. The DRM and the IRM improve the STOI scores, and the average improvement of the DRM over the IRM is approximately 0.021.

3.4. Evaluations with Real RIRs

Fig. 4 and Table 2 show the evaluation performance of the proposed approach and the IRM with the real impulse responses. For the STOI performance, the average STOI improvement of the DRM over the IRM is 0.20. When comparing with STOI at different RT60s (0.32 s, 0.47 s), the higher RT60 (0.47 s) causes worse separation performance, due to higher complexity. Besides, the direct to reverberant ratio (DDR) has positive effect on separation performance. For instance, by using the DRM, when the RT60 is 0.68, the performance is better than the one with lower RT60 (0.47 s), due to the influence of DDR, which strongly justifies another advantage of the geometric information based approach. PESQ performance is consistent with STOI performance.

In summary, the above experimental results confirm the proposed method can separate the target speech from the noisy reverberant mixture in both simulated and real room environments effectively. The proposed method outperforms the state-of-the-art method [2].

SNR Level		3 dB		0 dB		-3 dB	
RT60(s)	Targets	factory	babble	factory	babble	factory	babble
0.32	Unprocessed	1.02	1.25	0.74	0.99	0.56	0.78
	IRM	2.31	2.65	2.24	2.51	1.99	2.31
	DRM	**2.42**	**2.70**	**2.37**	**2.57**	**2.11**	**2.39**
0.47	Unprocessed	0.64	0.85	0.49	0.67	0.41	0.57
	IRM	2.17	2.43	1.99	2.31	1.80	2.14
	DRM	**2.28**	**2.53**	**2.11**	**2.40**	**1.89**	**2.21**
0.68	Unprocessed	0.74	0.91	0.69	0.80	0.52	0.61
	IRM	2.21	2.49	2.00	2.24	1.79	2.13
	DRM	**2.33**	**2.51**	**2.11**	**2.42**	**1.92**	**2.22**

Table 2: Averaged PESQ scores of 120 experiments for the IRM [2] and the proposed DRM systems at 3 dB, 0 dB and -3 dB SNR levels. The noisy reverberant speech mixtures are obtained by using the IEEE corpus and the factory and the babble background noise under real recorded impulse responses. The bold numbers represent the best performance.

4. CONCLUSIONS AND FUTURE WORK

We exploited the geometric information to provide the position of the target speaker and microphone to estimate the direct path impulse response, which is used to calculate the direct path speech. Based on the direct path speech, we calculated the DRM that is a new training target. The experimental results confirmed the DRM outperforms the state-of-the-art method.

In this study, the speaker position is physical stationary. For future research, more effort will be dedicated to improve the current proposed method for DNN based monaural speech separation for a moving source.

5. REFERENCES

[1] J. Li, R. Haeb-Umbach, and Y. Gong, *Robust Automatic Speech Recongnition: A Bridge to Practical Application*. Academic Press, 2015.

[2] Y. X. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[3] B. Rivet, W. Wang, S. M. Naqvi, and J. A. Chambers, "Audiovisual speech source separation: An overview of key methodologies," *IEEE Signal Processing Magazine*, vol. 19, no. 7, pp. 2125 – 2136, 2011.

[4] Y. Liang, J. Harris, S. M. Naqvi, G. Chen, and J. A. Chambers, "Independent vector analysis with a generalized multivariate Gaussian source prior for frequency domain blind source separation," *Signal Processing*, vol. 105, pp. 175–184, 2014.

[5] Z. Y. Zohny, S. M. Naqvi, and J. A. Chambers, "Enhancing MESSL algorithm with robust clustering based on student's t-distribution," *Electronics Letters*, vol. 50, pp. 552–554, 2014.

[6] Y. Sun, W. Rafique, J. A. Chambers, and S. M. Naqvi, "Underdetermined source separation using time-frequency masks and an adaptive combined Gaussian-Student's t probabilistic model," *Proc. of ICASSP*, 2017.

[7] Y. Sun, Y. Xian, P. Feng, J. A. Chambers, and S. M. Naqvi, "Estimation of the number of sources in measured speech mixtures with collapsed Gibbs sampling," *Proc. of SSPD*, 2017.

[8] Y. Jiang, D. L. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2112–2121, 2014.

[9] Y. Sun, L. Zhu, J. A. Chambers, and S. M. Naqvi, "Monaural source separation based on adaptive discriminative criterion in neural networks," in *Proc. of DSP*, 2017.

[10] A. Narayanan and D. L. Wang, "Ideal ratio mask estimation using deep neural network for robust speech separation," *Proc. of ICASSP*, 2013.

[11] D. S. Williamson, Y. X. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 17, pp. 483–492, 2016.

[12] P. Feng, W. Wang, S. Dlay, S. M. Naqvi, and J. A. Chambers, "Social force model-based MCMC-OCSVM particle phd filter for multiple human tracking," *IEEE Transactions on Multimedia*, vol. 19, no. 4, pp. 725–739, 2017.

[13] A. Rhemen, S. M. Naqvi, L. Mihaylova, and J. A. Chambers, "Multi-target tracking and occlusion handling with learned variational Bayesian clusters and a social force model," *IEEE Transactions on Signal Processing*, vol. 64, no. 5, pp. 1320–1335, 2016.

[14] M. S. Salman, S. M. Naqvi, A. Rehman, W. Wang, and J. A. Chambers, "Video-aided model-based source separation in real reverberant rooms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1900–1912, 2013.

[15] S. M. Naqvi, M. Yu, and J. A. Chambers, "A multimodal approach to blind source separation of moving sources," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, pp. 895–910, 2010.

[16] "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust*, no. 17, pp. 225–246, 1969.

[17] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun*, no. 12, pp. 247–251, 1993.

[18] J. B. Allen and D. A. Berkley, "Image method for efficiently simulation small-room acoustics," *Acoustical Society of America*, vol. 65, pp. 943–950, 1979.

[19] B. Shinn-Cunningham, N. Kopco, and T. Martin, "Localizing nearby sound sources in a classroom: Binaural room impulse responses," *Journal of the Acoustical Society of America*, vol. 117, pp. 3100–3115, 2005.

[20] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 31, no. 3, pp. 125–134, 2014.

[21] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[22] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *Proc. of ICASSP*, 1983.