A FEATURE COMPENSATION APPROACH USING PIECEWISE LINEAR APPROXIMATION OF AN EXPLICIT DISTORTION MODEL FOR NOISY SPEECH RECOGNITION

Jun Du[1]*, Qiang Huo*[2]

[1]University of Science and Technology of China, Hefei, Anhui, P. R. China
[2]Microsoft Research Asia, Beijing, P. R. China
(E-mails: unuedjwj@ustc.edu, qianghuo@microsoft.com)

ABSTRACT

This paper presents a new feature compensation approach to noisy speech recognition by using piecewise linear approximation (PLA) of an explicit model of environmental distortions. Two traditional approaches, namely vector Taylor series (VTS) and MAX approximations, are two special cases of our proposed approach. Formulations for maximum likelihood (ML) estimation of noise model parameters and minimum mean square error (MMSE) estimation of clean speech are derived. A hybrid approach of using different approximations for different types of noisy speech segments is also proposed. Experimental results on Aurora2 and Aurora3 databases demonstrate that the proposed approaches achieve consistently significant improvements in recognition accuracy compared to the traditional VTS-based feature compensation approach.

Index Terms— Robust speech recognition, feature compensation, piecewise linear approximation, distortion model.

1. INTRODUCTION

Most of current automatic speech recognition (ASR) systems use MFCCs (Mel-Frequency Cepstral Coefficients) and their derivatives as speech features, and a set of Gaussian mixture continuous density HMMs (CDHMMs) for modeling basic speech units. It is well-known that the performance of such an ASR system trained with clean speech will degrade significantly when the testing speech is distorted by additive noises. How to achieve the noise robustness has been an important research topic in ASR field. Among many approaches proposed previously, one type of approach is the so-called feature compensation approach using an *explicit* model of environmental distortions (e.g., [5, 4]). It is also the topic of this paper. For our approach, it is assumed that in the time domain, the "corrupted" speech $y[t]$ is subject to the following *explicit* distortion model:

$$y[t] = x[t] + n[t] \tag{1}$$

where independent signals $x[t]$ and $n[t]$ represent the t^{th} sample of clean speech and additive noise, respectively. By ignoring correlations among different filter banks, the distortion model in the log-power-spectral domain can be expressed *approximately* as

$$\exp(\mathbf{y}) = \exp(\mathbf{x}) + \exp(\mathbf{n}) \tag{2}$$

where $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{n}$ are log-power spectra in a particular channel of the filterbank of noisy speech, clean speech and noise, respectively. The nonlinear nature of the above distortion model makes statistical modeling and inference of the above variables difficult, therefore certain approximations have to be made. Understandably, two simple approximations have been tried in the past, namely linear approximation (*aka* the first-order vector Taylor series (VTS) approximation, e.g., [5, 4, 3]) and the so-called MAX approximation (e.g., [6, 8, 7]). In this paper, we propose a more accurate approximation approach by using a piecewise linear approximation (PLA) of the above nonlinear distortion model. To demonstrate its potential, as a first step, we propose and study a new feature compensation approach to robust noisy ASR in this paper.

The rest of the paper is organized as follows. In Section 2, we give an overview of the general formulation of our feature compensation approach. In Section 3, we present the detailed PLA formulation. In Section 4, we report some illustrative experimental results, and finally we conclude the paper in Section 5.

2. FEATURE COMPENSATION APPROACH

The flowchart of our feature compensation approach is illustrated in Fig. 1. In the training stage, a Gaussian mixture model (GMM) is trained from clean speech using MFCC features without cepstral mean normalization (CMN). Let's use $\{\omega_m, m = 1, 2, \cdots, M\}$ to denote the set of M mixture coefficient weights. In the recognition stage, first we transform the features and the clean-speech GMM from cepstral domain to the log-power-spectral domain by using the approach (i.e., IDCT) described in [3]. By ignoring the correlations among different channels of the filterbank, we can do feature compensation in the log-power-spectral domain for different channels independently.

Let's assume the noise feature $\mathbf{n}$ in this domain follows a Gaussian PDF (probability density function) with mean μ_n and variance σ_n^2 respectively. We have studied two ways of estimating $\{\mu_n, \sigma_n^2\}$. The first approach simply takes the sample mean and variance of the relevant features from the first several (10 in our experiments) frames of the unknown utterance. The second approach uses an ML estimation of $\{\mu_n, \sigma_n^2\}$ from the whole noisy speech utterance with T frames of observations, which can be solved by using EM algorithm iteratively (e.g., [7, 4]). The updating formulas are as follows:

$$\overline{\mu}_{n,d} = \frac{\sum_{t=1}^{T}\sum_{m=1}^{M} P(m|\mathbf{y}_t) E_n[n_{t,d}|y_{t,d}, m]}{\sum_{t=1}^{T}\sum_{m=1}^{M} P(m|\mathbf{y}_t)} \tag{3}$$

$$\overline{\sigma}_{n,d}^2 = \frac{\sum_{t=1}^{T}\sum_{m=1}^{M} P(m|\mathbf{y}_t) E_n[n_{t,d}^2|y_{t,d}, m]}{\sum_{t=1}^{T}\sum_{m=1}^{M} P(m|\mathbf{y}_t)} - \overline{\mu}_{n,d}^2 \tag{4}$$

This work was done while authors worked at The University of Hong Kong (HKU) and MSRA, and was partially supported by a grant from HKU's Seed Funding Programme for Basic Research.

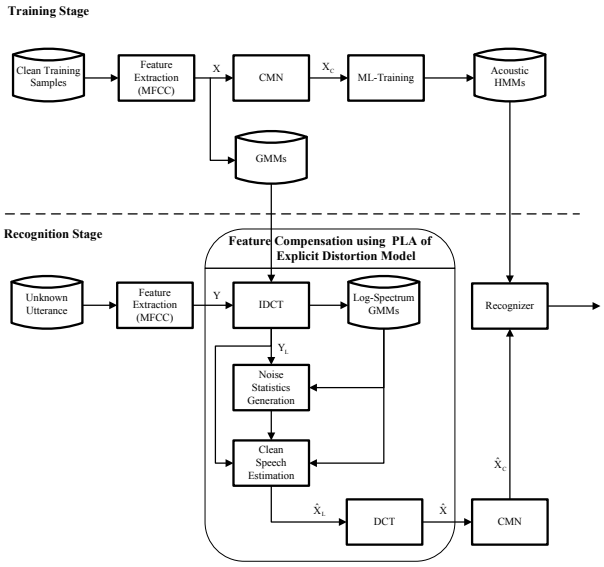

Fig. 1. Flowchart of our feature compensation approach.

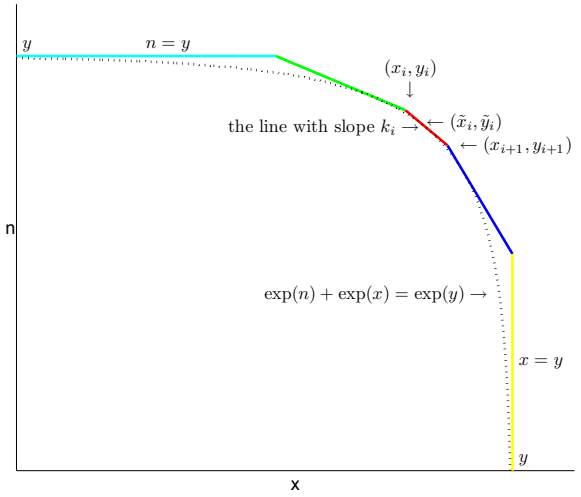

Fig. 2. Illustration of the piecewise linear approximation (PLA) of the explicit distortion model.

where

$$P(m|\mathbf{y}_t) = \frac{\omega_m p_y(\mathbf{y}_t|m)}{\sum_{l=1}^{M} \omega_l p_y(\mathbf{y}_t|l)} \ . \tag{5}$$

In the above equations, $p_y(\mathbf{y}_t|m)$ is the PDF of the noisy speech $\mathbf{y}_t$ for the m^{th} component of the compensated noisy speech mixture of densities, $E_n[n_{t,d}|y_{t,d}, m]$ and $E_n[n_{t,d}^2|y_{t,d}, m]$ are the relevant conditional expectations, t is the frame index, and d is the channel index of the filterbank.

Given the noisy speech and noise estimation, the minimum mean square error (MMSE) estimation of clean speech can be calculated as

$$\hat{\mathbf{x}}_t = E_x\left[\mathbf{x}_t|\mathbf{y}_t\right] = \sum_{m=1}^{M} P(m|\mathbf{y}_t) E_x\left[\mathbf{x}_t|\mathbf{y}_t, m\right] \tag{6}$$

where $E_x\left[\mathbf{x}_t|\mathbf{y}_t, m\right]$ is the conditional expectation of $\mathbf{x}_t$ given $\mathbf{y}_t$ for the m^{th} mixture component. Finally, the estimated clean-speech features in the log-power-spectral domain are transformed back to the cepstral domain using discrete cosine transform (DCT). The other modules in Fig. 1 are self-explained.

To implement the above feature compensation approach, the key technical issues become how to calculate $p_y(\mathbf{y}_t|m)$, $E_x\left[\mathbf{x}_t|\mathbf{y}_t, m\right]$, $E_x\left[\mathbf{x}_t^2|\mathbf{y}_t, m\right]$, $E_n[n_{t,d}|y_{t,d}, m]$, and $E_n[n_{t,d}^2|y_{t,d}, m]$, respectively. In next section, we elaborate on how the above problems can be solved if a PLA approximation of the nonlinear distortion function in Eq. (2) is used. For notational convenience, we drop hereinafter the indices related to the frame number, mixture component, and channel index of the filterbank without causing confusions.

3. PIECEWISE LINEAR APPROXIMATION (PLA) OF THE EXPLICIT DISTORTION MODEL

As illustrated in Fig. 2, in this paper, we propose to use a piecewise linear approximation (PLA) of the explicit model in Eq. (2) to characterize the relationship among $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{n}$. In the n-x plane, the curve representing the explicit model $y = f(x, n) = \log(\exp(x) + \exp(n))$ is approximated by N straight lines which are tangent to the curve. The slopes of these N lines, $\{k_i; i = 1, 2, \cdots, N\}$, have the property, $0 >= k_1 > ... > k_N > -\infty$. Obviously, as N increases, the PLA can be made increasingly more accurate.

For the i-th line with slope k_i, let's use (x_i, n_i), (x_{i+1}, n_{i+1}), and $(\tilde{x}_i, \tilde{n}_i)$ to denote the left endpoint, right endpoint, and the tangent point respectively. Then, the tangent point $(\tilde{x}_i, \tilde{n}_i)$ can be solved as:

$$\begin{cases} \tilde{x}_i = y + \ln \frac{-k_i}{1-k_i} \\ \tilde{n}_i = y + \ln \frac{1}{1-k_i} \end{cases} \ . \tag{7}$$

The equation of the i-th line is

$$n = k_i(x - \tilde{x}_i) + \tilde{n}_i = k_i x + (1 - k_i)y + b_i \triangleq g_i(x, y) \tag{8}$$

where

$$b_i = \ln \frac{1}{1-k_i} - k_i \ln \frac{-k_i}{1-k_i} \ . \tag{9}$$

The left endpoint (x_i, n_i) can be solved as:

$$\begin{cases} x_i = y + \triangle x_i \\ n_i = y + \triangle n_i \end{cases} \tag{10}$$

where

$$\begin{cases} \triangle x_i = \frac{b_{i-1} - b_i}{k_i - k_{i-1}} \\ \triangle n_i = \frac{k_i b_{i-1} - k_{i-1} b_i}{k_i - k_{i-1}} \end{cases} \ . \tag{11}$$

Given the following Gaussian PDFs of x and n,

$$p_x(x) = \mathcal{N}(x, \mu_x, \sigma_x) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left[-\frac{(x - \mu_x)^2}{2\sigma_x^2}\right], \tag{12}$$

$$p_n(n) = \mathcal{N}(n, \mu_n, \sigma_n) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{(n - \mu_n)^2}{2\sigma_n^2}\right], \tag{13}$$

the corresponding cumulative density function (CDF) $c_x(x)$ and $c_n(n)$ can be calculated as

$$c_x(x) = \int_{-\infty}^{x} p_x(u)du = \Phi(x, \mu_x, \sigma_x) \tag{14}$$

$$c_n(n) = \int_{-\infty}^{n} p_n(v)dv = \Phi(n, \mu_n, \sigma_n) \tag{15}$$

where $\Phi(x, \mu_x, \sigma_x)$ and $\Phi(n, \mu_n, \sigma_n)$ can be obtained by lookup tables.

Given the above notations and results, we present in the following subsections the main results that are required in implementing our feature compensation approach.

3.1. Calculating $p_y(y)$

$$p_y(y) = \oint_{f(x,n)=y} p_{xn}(x,n) = \frac{d}{dy}\left[\iint_{f(x,n)<=y} p_{xn}(x,n)dndx\right] \quad (16)$$

where $p_{xn}(x,n)$ is the joint distribution of x and n. By using PLA, we first define

$$\iint_{f(x,n)<=y} \triangleq \sum_{i=1}^{N} \int_{x_i}^{x_{i+1}} \int_{-\infty}^{g_i(x,y)} \quad (17)$$

then we have:

$$p_y(y) = \sum_{i=1}^{N} p_{y,i}(y)\left|c_{y,i,i+1}(y) - c_{y,i,i}(y)\right| \quad (18)$$

where

$$p_{y,i}(y) = \mathcal{N}(y, \mu_{y,i}, \sigma_{y,i}) \quad (19)$$
$$c_{y,i,i}(y) = \Phi(y, \mu_{y,i,i}, \sigma_{y,i,i}) \quad (20)$$
$$c_{y,i,i+1}(y) = \Phi(y, \mu_{y,i,i+1}, \sigma_{y,i,i}) \quad (21)$$

and

$$\mu_{y,i} = \frac{\mu_n - b_i - k_i\mu_x}{1-k_i} \quad (22)$$
$$\sigma_{y,i} = \frac{\sqrt{\sigma_n^2 + k_i^2\sigma_x^2}}{1-k_i} \quad (23)$$
$$\mu_{y,i,i} = \frac{\sigma_n^2(\mu_x - \triangle x_i) + k_i\sigma_x^2(\mu_n - \triangle n_i)}{\sigma_n^2 + k_i\sigma_x^2} \quad (24)$$
$$\sigma_{y,i,i} = \frac{\sigma_x\sigma_n\sqrt{\sigma_n^2 + k_i^2\sigma_x^2}}{|\sigma_n^2 + k_i\sigma_x^2|} \quad (25)$$
$$\mu_{y,i,i+1} = \frac{\sigma_n^2(\mu_x - \triangle x_{i+1}) + k_i\sigma_x^2(\mu_n - \triangle n_{i+1})}{\sigma_n^2 + k_i\sigma_x^2}. \quad (26)$$

3.2. Calculating $E_x[x|y]$ and $E_x[x^2|y]$

$$E_x[x|y] = \frac{1}{p_y(y)} \oint_{f(x,n)=y} x p_{xn}(x,n) = \frac{1}{p_y(y)}\left[\frac{d}{dy}\iint_{f(x,n)<=y} x p_{xn}(x,n)dndx\right] \quad (27)$$

By using PLA, we have:

$$E_x[x|y] = \frac{1}{p_y(y)}\sum_{i=1}^{N} p_{y,i}(y)\left(\left|c_{y,i,i+1}(y) - c_{y,i,i}(y)\right|\mu_{x,i} - \sigma_{x,i}\sigma_{y,i,i}\left[p_{y,i,i+1}(y) - p_{y,i,i}(y)\right]\right) \quad (28)$$

where

$$p_{y,i,i}(y) = \mathcal{N}(y, \mu_{y,i,i}, \sigma_{y,i,i}) \quad (29)$$
$$p_{y,i,i+1}(y) = \mathcal{N}(y, \mu_{y,i,i+1}, \sigma_{y,i,i+1}) \quad (30)$$

and

$$\mu_{x,i} = \frac{\sigma_n^2\mu_x + k_i^2\sigma_x^2\bar{\mu}_{x,i}}{\sigma_n^2 + k_i^2\sigma_x^2} \quad (31)$$
$$\bar{\mu}_{x,i} = \frac{\mu_n - b_i - (1-k_i)y}{k_i} \quad (32)$$
$$\sigma_{x,i} = \frac{\sigma_x\sigma_n}{\sqrt{\sigma_n^2 + k_i^2\sigma_x^2}}. \quad (33)$$

$$E_x[x^2|y] = \frac{1}{p_y(y)} \oint_{f(x,n)=y} x^2 p_{xn}(x,n) = \frac{1}{p_y(y)}\left[\frac{d}{dy}\iint_{f(x,n)<=y} x^2 p_{xn}(x,n)dndx\right] \quad (34)$$

By using PLA, we have:

$$E_x[x^2|y] = \frac{1}{p_y(y)}\sum_{i=1}^{N} p_{y,i}(y)(\left|c_{y,i,i+1}(y) - c_{y,i,i}(y)\right| (\mu_{x,i}^2 + \sigma_{x,i}^2) - \sigma_{x,i}\sigma_{y,i,i}\left[(\mu_{x,i} + x_{i+1}) p_{y,i,i+1}(y) - (\mu_{x,i} + x_i)p_{y,i,i}(y)\right]). \quad (35)$$

3.3. Calculating $E_n[n|y]$ and $E_n[n^2|y]$

Similar to the derivation in section 3.2, we can have

$$E_n[n|y] = \frac{1}{p_y(y)}\sum_{i=1}^{N} p_{y,i}(y)\left(\left|c_{y,i,i+1}(y) - c_{y,i,i}(y)\right|\mu_{n,i} + \sigma_{n,i}\sigma_{y,i,i}\left[p_{y,i,i+1}(y) - p_{y,i,i}(y)\right]\right) \quad (36)$$

$$E_n[n^2|y] = \frac{1}{p_y(y)}\sum_{i=1}^{N} p_{y,i}(y)(\left|c_{y,i,i+1}(y) - c_{y,i,i}(y)\right| (\mu_{n,i}^2 + \sigma_{n,i}^2) + \sigma_{n,i}\sigma_{y,i,i}\left[(\mu_{n,i} + n_{i+1}) p_{y,i,i+1}(y) - (\mu_{n,i} + n_i)p_{y,i,i}(y)\right]) \quad (37)$$

where

$$\mu_{n,i} = \frac{\sigma_n^2\bar{\mu}_{n,i} + k_i^2\sigma_x^2\mu_n}{\sigma_n^2 + k_i^2\sigma_x^2} \quad (38)$$
$$\bar{\mu}_{n,i} = k_i\mu_x + (1-k_i)y + b_i \quad (39)$$
$$\sigma_{n,i} = \frac{|k_i|\sigma_x\sigma_n}{\sqrt{\sigma_n^2 + k_i^2\sigma_x^2}}. \quad (40)$$

3.4. Discussions

It is interesting to note the following facts:

- When $N = 1$, PLA with $k_1 = -\exp(x_0 - n_0)$ becomes equivalent to the first-order VTS approximation with (x_0, n_0) as the expansion point [5];
- When $N = 2$, PLA with $k_1 = 0$ and $k_2 = -\infty$ becomes MAX approximation [6];

Table 1. Performance (word accuracy in %) comparison of several feature compensation approaches using different noise estimation methods, averaged over SNRs between 0 and 20 dB across all noise conditions on three different test sets (i.e., Sets A, B, and C) of Aurora2 database. The baseline performance is 68.74%.

Methods of Noise Reestimation	Methods of Feature Compensation			
	VTS	MAX	PLA(3)	MAX/PLA(3)
No Reestimation	77.78	81.53	80.56	82.10
VTS-based	84.02	84.53	84.71	85.20
MAX-based	84.60	84.79	85.19	85.52

- When $N = 3$, PLA with $k_1 = 0$, $k_2 = -\exp(x_0 - n_0)$ and $k_3 = -\infty$, referred to as PLA(3) hereinafter, offers a more accurate model than both VTS and MAX approximation for contour integration.

In the past, only one specific approximation is used for the compensation of all frames in an unknown utterance. However, it is well-known that the MAX approximation is quite accurate for the cases of either very low and very high SNRs. This motivates us to propose the following hybrid approach:

- For $y < \mu_n$, we use MAX approximation; Otherwise, we use PLA(3).

In the following, we use "MAX/PLA(3)" to refer to the above hybrid approach.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

In order to verify the effectiveness of the proposed method, a series of experiments are performed for the task of recognition of connected digit strings on both Aurora2 and Aurora3 (Finnish subset) databases. A full description of the two databases is given in [1, 2].

In our ASR systems, the feature vector we used consists of 13 MFCCs (including C_0) plus their first and second order derivatives. The number of Mel-frequency filter banks is 23. The cepstra are computed based on the power spectra. Each digit is modeled by a whole word left-to-right CDHMM, which consists of 16 emitting states, each having 3 Gaussian mixture components. The mixture number of clean-speech GMM for feature compensation is 256. For Aurora2 database, "clean-training" is used. For Aurora3 Finnish database, we only focus on high-mismatch (HM) condition. In HM condition, training data includes utterances recorded by close-talking (CT) microphone, which can be considered as "clean", while testing data is recorded by hands-free (HF) microphone. Our baseline systems refer to the ones with CMN but no other feature compensation applied.

4.2. Experimental Results

Tables 1 and 2 summarize a performance (word accuracy in %) comparison among several feature compensation approaches for cases of using first 10 frames to estimate noise model parameters (referred to as "No Reestimation"), and using VTS-based or MAX-based methods (seven EM iterations in both cases) for noise reestimation. Performances of respective baseline systems are also included for comparison. It is observed that 1) Compared with the baseline performance, significant improvements are achieved by all the feature compensation methods; 2) Without noise reestimation, the performance of "MAX/PLA(3)" is much better than that of VTS; 3) Given the same noise estimation, "MAX/PLA(3)" always achieves the best performance; 4) In terms of noise estimation, MAX model outperforms VTS.

Table 2. Performance (word accuracy in %) comparison of several feature compensation approaches using different noise estimation methods in the high-mismatch (HM) condition on Aurora3 Finnish database. The baseline performance is 76.22%.

Methods of Noise Reestimation	Methods of Feature Compensation			
	VTS	MAX	PLA(3)	MAX/PLA(3)
No Reestimation	77.77	80.32	80.99	83.60
VTS-based	83.67	83.57	84.45	85.44
MAX-based	84.03	84.35	84.66	86.08

5. CONCLUSION AND FUTURE WORKS

In this paper, we have proposed a feature compensation approach using piecewise linear approximation (PLA) of an explicit distortion model and verified its effectiveness on both Aurora2 and Aurora3 tasks. Ongoing and future works include 1) to study PLA-based HMM compensation, 2) to explore irrelevant variability normalization (IVN) based HMM training using PLA, 3) to apply the similar idea to speech enhancement, 4) to investigate how to extend the PLA-based formulation such that correlations among different filter banks can also be considered. We will report those results elsewhere when they become available.

6. REFERENCES

[1] Aurora document AU/217/99, "Availability of Finnish speechdat-car database for ETSI STQ WI008 front-end standardisation," Nokia, 1999.

[2] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," *Proc. ISCA ITRW ASR*, 2000, pp.181-188.

[3] Y. Hu and Q. Huo, "Irrelevant variability normalization based HMM training using VTS approximation of an explicit model of environmental distortions," *Proc. Interspeech*, 2007, pp.1042-1045.

[4] D.-Y. Kim, C.-K. Un, and N.-S. Kim, "Speech recognition in noisy environments using first-order vector Taylor series," *Speech Communication*, Vol. 24, pp.39-49, 1998.

[5] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," *Proc. ICASSP*, 1996, pp.733-736.

[6] A. N*á*das, D. Nahamoo, and M. A. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 37, No. 10, pp. 1495-1503, 1989.

[7] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, "Integrated models of signal and background with application to speaker identification in noise," *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 2, pp.245-257, 1994.

[8] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," *Proc. ICASSP*, 1990, pp.845-848.