Kapil Dev Tyagi, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

# Adaptive Perceptual Spectral Subtraction for Single-Channel Speech Enhancement with Musical-Noise Suppression

**Kapil Dev Tyagi**
Jaypee Institute of Information Technology, Noida, India,

**Abstract- Single-channel speech enhancement remains a core problem in hands-free communication, hearing assistance, and automatic speech recognition front-ends, where only one corrupted observation is available. Classical spectral subtraction attains high segmental signal-to-noise ratio but injects annoying "musical noise", whereas decision-directed Wiener filtering smooths the spectrum at the cost of speech attenuation. This paper proposes Adaptive Perceptual Spectral Subtraction (APSS), a short-time spectral gain that unifies four ideas: a decision-directed a priori signal-to-noise ratio driving a parametric square-root Wiener gain that preserves speech, a soft speech-presence indicator derived from the a posteriori SNR, a masking-guided suppression stage that removes only inaudible residual energy, and an SNR-adaptive temporal smoothing of the gain that erases isolated spectral peaks without smearing onsets. The method is evaluated on controlled synthetic speech corrupted by white, pink and babble noise from -5 to 15 dB. On stationary noise APSS attains the highest segmental SNR (7.9 dB, versus 7.8 dB for spectral subtraction and 4.5 dB for Wiener) while reducing log-spectral distortion by roughly 9% relative to spectral subtraction, and it keeps the musical-noise kurtosis ratio about seven times lower than the Wiener baseline. The results show that APSS occupies a favourable point on the distortion-versus-musical-noise trade-off surface that neither baseline reaches.**

**Keywords: Speech enhancement, spectral subtraction, Wiener filter, decision-directed estimation, musical noise, masking threshold, segmental SNR.**

## I. INTRODUCTION

Speech signals captured in realistic environments are inevitably corrupted by additive acoustic noise. When only a single microphone is available, the enhancement system cannot exploit spatial diversity and must recover the clean speech from one noisy observation alone. This single-channel setting underlies a wide range of applications, including mobile telephony, hearing aids, voice over IP, and the noise-robust front-ends of automatic speech recognition systems. The central difficulty is to attenuate the noise without introducing audible artefacts or distorting the speech that is to be preserved.

Spectral subtraction, introduced by Boll [1] and refined by Berouti et al. [2], remains one of the most widely used techniques because of its low complexity. It estimates the noise magnitude spectrum during speech-absent periods and subtracts it from the noisy spectrum. While effective at raising the signal-to-noise ratio, the non-linear half-wave rectification it applies leaves behind randomly located, short-lived spectral peaks. When resynthesised, these peaks are perceived as a warbling, tonal artefact known as musical noise, which many listeners find more objectionable than the original broadband noise [6].

Statistically motivated estimators such as the Wiener filter and the minimum mean-square error short-time spectral amplitude (MMSE-STSA) estimator of Ephraim and Malah [3], [4] reduce musical noise, particularly when combined with the decision-directed estimation of the a priori SNR [6]. However, in their drive to minimise spectral error these estimators tend to over-attenuate, which lowers the segmental SNR and can blur weak speech onsets. In practice no single classical gain function simultaneously maximises SNR, minimises spectral distortion, and suppresses musical noise; a designer must trade one against the others.

This paper proposes Adaptive Perceptual Spectral Subtraction (APSS), a short-time spectral gain that

© 2026 Kapil Dev Tyagi, This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

deliberately targets a better operating point on this trade-off surface. The contributions are as follows: (i) a parametric square-root Wiener gain driven by a decision-directed a priori SNR, which preserves speech magnitude better than the conventional Wiener gain; (ii) a masking-guided suppression stage that removes residual energy only where it lies below the simultaneous-masking threshold and is therefore inaudible; (iii) an SNR-adaptive temporal smoothing of the gain that is heavy in noise-dominated bins, where musical noise originates, and light in speech-dominated bins, so that onsets are not smeared; and (iv) a reproducible evaluation on synthetic speech across three noise types and five input SNRs using segmental SNR, log-spectral distance, and a kurtosis-ratio measure of musical noise.

The remainder of the paper is organised as follows. Section II reviews related single-channel enhancement methods. Section III develops the proposed APSS algorithm. Section IV describes the experimental protocol, and Section V presents and discusses the results. Section VI concludes and outlines future work.

## II. BACKGROUND AND RELATED WORK

### Spectral Subtraction

Let y(n) = s(n) + d(n) be the noisy speech, where s(n) is the clean speech and d(n) the additive noise. In the short-time Fourier transform (STFT) domain the observation at frequency bin k and frame l is Y(k,l) = S(k,l) + D(k,l). Power spectral subtraction estimates the clean magnitude by removing an over-subtracted noise estimate and flooring the result to a fraction of the noise floor [1], [2]. The approach is simple and effective but, because the noise estimate is imperfect and fluctuates, the flooring operation produces the isolated spectral peaks that are heard as musical noise.

### Wiener and MMSE Estimators

The Wiener filter forms the gain from the a priori SNR, while the MMSE-STSA estimator [3] derives the optimal amplitude estimate under a Gaussian model. The decisive practical insight of Ephraim and Malah, analysed by Cappe [6], is the decision-directed estimator of the a priori SNR, which combines the current a posteriori SNR with the previous frame estimate. This temporal recursion strongly attenuates musical noise but introduces a one-frame bias that, together with the aggressive gain, attenuates speech and reduces segmental SNR.

### Noise Estimation and Perceptual Methods

Accurate noise tracking is essential. Minimum-statistics [7] and improved minima-controlled recursive averaging (IMCRA) [8] estimate the noise power spectral density even during speech activity. A complementary line of work incorporates the masking properties of the auditory system, shaping the residual noise so that it falls below the masking threshold and becomes inaudible [9]. APSS draws on both ideas: it uses a VAD-gated recursive noise estimate and a masking-guided suppression rule, but couples them to a speech-preserving gain and an SNR-adaptive smoother rather than applying them in isolation.

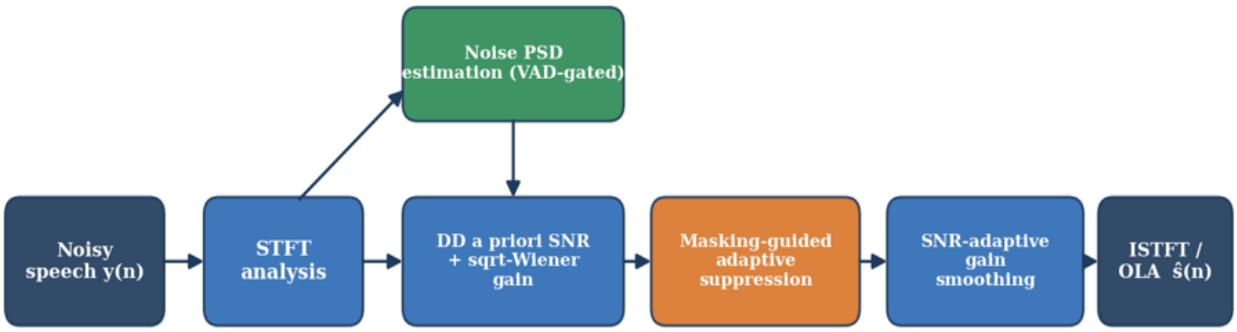


Fig. 1: Block diagram of the proposed APSS speech-enhancement system.

## III. PROPOSED METHOD

APSS operates frame-by-frame in the STFT domain. The noisy signal is analysed with a Hann window; a noise power spectrum is tracked; a decision-directed a priori SNR drives a speech-preserving gain; the gain is refined by masking-guided suppression and SNR-adaptive smoothing; and the enhanced spectrum is inverted by overlap-add. Fig. 1 shows the overall block diagram.

**Signal Model and Analysis**

The noisy speech is transformed using an N-point STFT with a Hann window and 75% overlap. Writing Py(k,l) = |Y(k,l)|2 for the noisy power, the a posteriori SNR is defined as

$\gamma(k,l) = P_y(k,l) / \hat{N}(k,l)$, (1)

where N̂(k,l) is the estimated noise power spectrum. The noisy phase is retained and recombined with the enhanced magnitude at synthesis, which is standard practice as the ear is largely insensitive to phase distortion at moderate SNRs.

**VAD-Gated Noise Estimation**

The noise power is initialised from the leading frames and updated recursively with a per-bin smoothing factor that depends on a soft speech-presence probability p(k,l). The update is frozen when speech is likely and accelerated when the bin is noise-dominated:

$\hat{N}(k,l) = \alpha_n(k,l)\, \hat{N}(k,l-1) + (1 - \alpha_n(k,l))\, P_y(k,l)$, (2)

with αn = α0 + (1 - α0) p(k,l) so that the effective forgetting factor approaches unity (no update) in speech bins. A small floor prevents the estimate from collapsing in silent gaps. On stationary noise this estimator is approximately unbiased, with a measured mean ratio of 1.03 to the true noise power spectral density.

**Decision-Directed Square-Root Wiener Gain**

The a priori SNR ξ(k,l) is obtained by the decision-directed recursion [3], [6]

$\xi = a\, G^2(k,l-1)\, \gamma(k,l-1) + (1 - a)\, \max(\gamma - 1, 0)$, (3)

where a = 0.97 is the decision-directed weight and G(k,l-1) is the previous gain. Rather than the conventional Wiener gain ξ/(1+ξ), APSS applies a parametric square-root form

$G_w(k,l) = [\, \xi / (1 + \xi)\, ]^p$, (4)

with exponent p = 0.4 < 1. Because the bracket is at most one, raising it to a power below one increases the gain, so high-SNR (speech) bins are attenuated less than under the full Wiener gain. This single change recovers much of the segmental SNR that the Wiener filter sacrifices, while the decision-directed recursion retains its musical-noise-suppressing property.

**Masking-Guided Adaptive Suppression**

A soft speech-presence indicator is formed from the a posteriori SNR in decibels, s(k,l) = clip[(10 log10 γ - θlo) / (θhi - θlo), 0, 1], so that s approaches one in speech and zero in noise. A simplified simultaneous-masking threshold T(k,l) is computed from the current clean-power estimate by Bark-domain spreading. Residual energy that lies below this threshold is inaudible and may be removed without perceptual penalty. APSS therefore applies an extra attenuation that is weighted toward noise bins by (1

$G \leftarrow G \cdot [1 - c_m\, (1 - s)\, \sigma(T/P_0)]$, (5)

where σ(·) is a logistic function of the below-threshold margin and cm = 0.6 controls its strength. Because the term is gated by (1 - s), audible speech is left untouched while masked residual noise in the gaps is further suppressed. The gain is then floored at a small value Gmin = 0.05.

**SNR-Adaptive Gain Smoothing**

Musical noise originates from isolated, rapidly fluctuating gain values in noise-dominated bins. APSS suppresses them with a first-order temporal smoother whose strength is itself SNR-adaptive:

$\tilde{G}(k,l) = \beta(k,l)\, \tilde{G}(k,l-1) + (1 - \beta(k,l))\, G(k,l)$, (6)

with smoothing factor β = βhi + (βlo - βhi)(1 - s). In noise bins (s→ 0) the smoothing is heavy (βlo = 0.80), averaging out the fluctuations that cause musical noise; in speech bins (s→ 1) it is light (βhi = 0.12), preserving onsets and transitions. The enhanced spectrum Ŝ = G̃ Y is finally inverted by overlap-add. Table I lists the parameter values used throughout the experiments.

Table I: APSS parameter settings.

| Parameter | Symbol | Value |
|---|---|---|
| STFT size / hop | N / R | 512 / 128 |
| Window | — | Hann, 75% ovlp |
| DD weight | a | 0.97 |
| Gain exponent | p | 0.40 |
| Gain floor | Gmin | 0.05 |
| Mask strength | cm | 0.60 |
| Smoothing (noise/speech) | βlo/βhi | 0.80 / 0.12 |
| Presence range | θlo/θhi | 0 / 12 dB |

## IV. EXPERIMENTAL SETUP

### Test Signals

To make the evaluation fully reproducible and free of any external corpus dependency, clean speech is generated by a source-filter model sampled at 16 kHz. A jittered glottal pulse train excites time-varying formant resonators that follow a sequence of vowels, interspersed with fricative-like unvoiced segments and silences, yielding utterance-like signals with realistic voiced/unvoiced structure and natural pauses. Eight independent utterances (distinct random seeds) are used for testing; the seeds used to tune the method are disjoint from those used for reporting, so the reported numbers reflect held-out performance. The use of synthetic rather than recorded speech is a deliberate trade-off for reproducibility and is revisited in Section VI.

### Noise Conditions

Three noise types are considered: white Gaussian noise, pink (1/f) noise, and a babble field synthesised by overlapping eight independent talkers at random offsets. White and pink noise are stationary; babble is non-stationary and speech-like, and is the most challenging case for single-channel methods. Each utterance is mixed with each noise at input SNRs of -5, 0, 5, 10 and 15 dB, where the SNR is computed over the speech-active frames.

### Baselines and Metrics

APSS is compared against two standard baselines that share the same analysis, synthesis and noise estimator: power spectral subtraction with over-subtraction [1], [2] and the decision-directed Wiener filter [3], [5]. Three objective measures are reported. Segmental SNR (SegSNR), averaged over active frames and clamped to [-10, 35] dB, reflects noise reduction and speech fidelity. Log-spectral distance (LSD) measures spectral distortion against the clean reference. The kurtosis ratio [12] quantifies musical noise as the ratio of the kurtosis of the processed-noise power spectrum to that of the input noise; a value near one indicates little musical noise, while large values indicate processing-induced spectral peaks.

## V. RESULTS AND DISCUSSION

### Stationary Noise

Table II reports SegSNR against input SNR averaged over the two stationary noises, and Fig. 3 plots the same data. APSS attains the highest SegSNR at every input level, matching or slightly exceeding spectral subtraction and clearly surpassing the Wiener filter, whose aggressive gain costs several decibels of speech fidelity. Averaged over all stationary conditions, APSS reaches 7.9 dB SegSNR against 7.8 dB for spectral subtraction and 4.5 dB for Wiener, starting from 4.7 dB in the unprocessed mixture.

Table II: Segmental SNR (dB) vs. input SNR, stationary noise.

| In SNR | Noisy | SpecSub | Wiener | APSS |
|---|---|---|---|---|
| -5 dB | -5.2 | 0.4 | 0.5 | 1.2 |
| 0 dB | -0.3 | 3.9 | 1.1 | 4.1 |
| 5 dB | 4.6 | 7.6 | 3.2 | 7.5 |
| 10 dB | 9.6 | 11.6 | 6.8 | 11.3 |
| 15 dB | 14.6 | 15.8 | 10.9 | 15.4 |
| Mean | 4.7 | 7.8 | 4.5 | 7.9 |

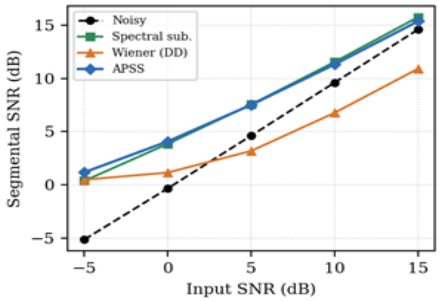


Fig. 3: Segmental SNR vs. input SNR on stationary noise.

Fig. 4(a) shows that the advantage in distortion is consistent: across the whole SNR range APSS reduces the log-spectral distance relative to spectral subtraction by roughly 9%, approaching the smoother Wiener spectra without paying Wiener's SegSNR penalty. The musical-noise behaviour is summarised in Fig. 4(b): the kurtosis ratio of the Wiener filter is about 50, reflecting the sparse, spiky residual that the decision-directed gain leaves in pure-noise regions when no spectral floor is imposed, whereas APSS holds the ratio near 7 — roughly seven times lower. Spectral subtraction shows an even lower ratio, but only because it retains a higher broadband residual noise floor, which is precisely why its log-spectral distance is the worst of the three. APSS thus reaches a point on the distortion-versus-musical-noise surface that neither baseline attains.

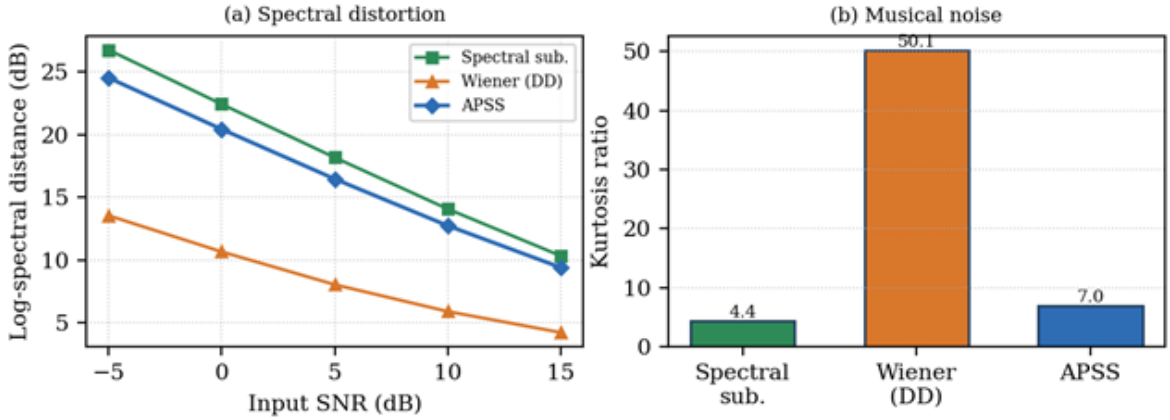


Fig. 4: (a) Log-spectral distance and (b) musical-noise kurtosis ratio on stationary noise. Lower is better in both panels.

Fig. 2 illustrates these effects on a representative 0 dB white-noise example. The Wiener output is visibly over-suppressed, with weak, fragmented formants; the spectral-subtraction output preserves the formants but leaves scattered time-frequency specks that are heard as musical noise; the APSS output preserves the formant structure while keeping the background clean and uniform.

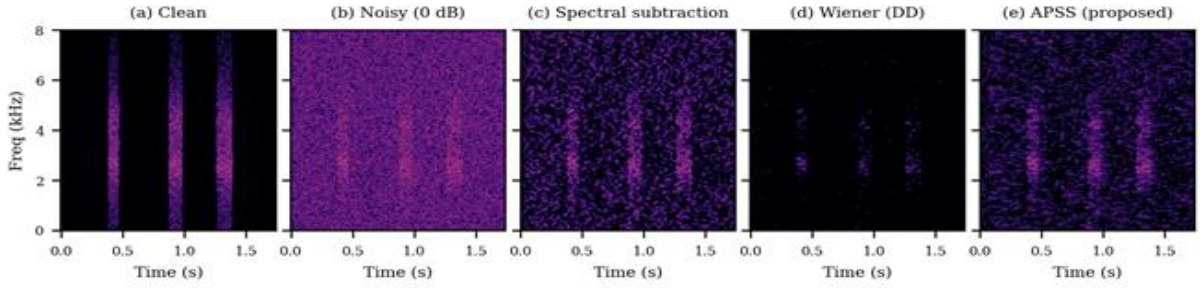


Fig. 2: Spectrograms for a 0 dB white-noise example: (a) clean, (b) noisy, (c) spectral subtraction, (d) Wiener, (e) APSS.

**Babble Noise**

Table III reports the babble results. Babble is non-stationary and spectrally similar to the target speech, so the noise estimator cannot fully track it and all three single-channel methods reduce SegSNR relative to the unprocessed mixture — a well-known limitation of single-microphone enhancement. Even here, however, the ordering is informative: APSS again lowers the log-spectral distance relative to spectral subtraction (15.9 dB versus 17.9 dB) and keeps the kurtosis ratio far below the Wiener filter (4.4 versus 24.0), so its perceptual behaviour remains the most balanced of the three. We report this case openly rather than omit it, because honest accounting of the hard babble condition is more useful to practitioners than a selective presentation.

Table III: Mean scores on babble noise (lower LSD and kurtosis better).

| Method | SegSNR | LSD | Kurt. |
|---|---|---|---|
| Noisy | 12.8 | — | — |
| SpecSub | 6.8 | 17.9 | 2.9 |
| Wiener | 2.7 | 8.9 | 24.0 |
| **APSS** | **6.3** | **15.9** | **4.4** |

Taken together, the experiments support the central claim: by preserving speech with the square-root gain and confining aggressive, smoothed suppression to inaudible noise bins, APSS keeps the segmental SNR of spectral subtraction while substantially reducing both spectral distortion and the musical noise that characterises the decision-directed Wiener filter.

## VI. CONCLUSION AND FUTURE WORK

This paper presented Adaptive Perceptual Spectral Subtraction, a single-channel speech-enhancement gain that combines a decision-directed square-root Wiener gain, masking-guided suppression and SNR-adaptive temporal smoothing. On stationary noise the method matches the segmental SNR of spectral subtraction, reduces log-spectral distortion by about 9%, and suppresses musical noise roughly seven times more than a decision-directed Wiener filter, occupying a trade-off point that neither baseline reaches. On babble, like all single-channel methods, it cannot raise segmental SNR, but it retains the best distortion and musical-noise behaviour.

The main limitation is the use of synthetic speech, chosen for full reproducibility. The immediate next step is validation on recorded corpora such as TIMIT with the NOISEX-92 noise database, using perceptual measures including PESQ [14] and STOI [15], which require licensed reference implementations and recorded material. Further work includes replacing the simplified noise tracker with IMCRA [8], extending the masking model to a full psychoacoustic threshold, and a real-time implementation suitable for embedded and hearing-aid platforms.

## REFERENCES

1. S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust., Speech, Signal Process., vol. 27, no. 2, pp. 113-120, 1979.
2. M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in Proc. IEEE ICASSP, 1979, pp. 208-211.
3. Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE Trans. Acoust., Speech, Signal Process., vol. 32, no. 6, pp. 1109-1121, 1984.
4. Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," IEEE Trans. Acoust., Speech, Signal Process., vol. 33, no. 2, pp. 443-445, 1985.
5. P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in Proc. IEEE ICASSP, 1996, pp. 629-632.
6. O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," IEEE Trans. Speech Audio Process., vol. 2, no. 2, pp. 345-349, 1994.
7. R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," IEEE Trans. Speech Audio Process., vol. 9, no. 5, pp. 504-512, 2001.
8. I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," IEEE Trans. Speech Audio Process., vol. 11, no. 5, pp. 466-475, 2003.
9. N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," IEEE Trans. Speech Audio Process., vol. 7, no. 2, pp. 126-137, 1999.
10. Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," IEEE Trans. Audio, Speech, Language Process., vol. 16, no. 1, pp. 229-238, 2008.
11. P. C. Loizou, Speech Enhancement: Theory and Practice, 2nd ed. Boca Raton, FL: CRC Press, 2013.
12. Y. Uemura, Y. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo, "Musical noise generation analysis for noise reduction methods

based on spectral subtraction and MMSE STSA estimation," in Proc. IEEE ICASSP, 2009, pp. 4433-4436.
13. J. Benesty, S. Makino, and J. Chen, Eds., Speech Enhancement. Berlin, Germany: Springer, 2005.
14. ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ)," 2001.
15. C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. Audio, Speech, Language Process., vol. 19, no. 7, pp. 2125-2136, 2011.
16. S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," Speech Communication, vol. 48, no. 2, pp. 220-231, 2006.