情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2011-SLP-87 No.14
2011/7/23

ウエーブレットに基づくウイナーフィルタを用いた雑音及び残響に頑健な音声認識

ゴメス・ランディ †1　　河原　達也†1

頑健な音声認識のために、ウエーブレット領域で雑音と残響を抑圧する方法を提案する。ウエーブレット変換のパラメータは、音声・背景雑音・遅い残響成分の各々に対して最適化し、効果的なウイナーフィルタを行うためのウイナーゲインを求める。具体的には、背景雑音と遅い残響成分を抑圧するためのウイナーゲインを独立に求めた後、両者を組み合わせる。様々な雑音や残響条件に対応できるように、雑音プロファイルと残響時間の自動同定を導入している。提案手法を大語彙連続音声認識において評価し、既存の手法に比べて有効性を確認した。

Robust Speech Recognition in Noisy and Reverberant Environments Using Wavelet-based Wiener Filtering

Randy Gomez †1 and Tatsuya Kawahara†1

We present a method of enhancing the speech signal corrupted by noise and late reflection in the wavelet domain for robust automatic speech recognition (ASR). The wavelet parameters for speech, background noise and late reflection are optimized to achieve a better estimate of the Wiener gain for effective filtering. Wiener gains to compensate for the effects of background noise and late reflection are independently estimated and then combined. To cope with different noise and reverberant conditions, we introduce the noise profiles and reverberation time identification. The proposed method is evaluated in a large vocabulary continuous speech recognition (LVCSR) task, and shown to outperform several conventional methods.

†1 京都大学 学術情報メディアセンター
Academic Center for Computing and Media Studies (ACCMS), Kyoto University.
Randy Gomez is a research fellow of the Japan Society for Promotion of Science (JSPS).

1. Introduction

In real-environment conditions, the speech signal is often contaminated with noise and reverberation resulting to mismatch in the acoustic model (AM). Thus, speech enhancement including denoising and dereverberation is one of the most important topics in ASR. While speech enhancement has been conventionally studied independently from ASR, we are studying on tight integration of enhancement and ASR using a maximum likelihood criterion[1].

The model of the reverberant speech $X(w,f)$ (short-term spectrum, w: window frame, f: frequency) we adopt is based on the additive effects of the early $X_E(w,f)$ and late $X_L(w,f)$ reflection,

$$\begin{aligned} X(w,f) &\approx \quad X_E(w,f) \quad + \quad X_L(w,f) \\ &\approx S(w,f)H(0,f) + \textstyle\sum_{d=1}^{D} S(w-d,f)H(d,f) \end{aligned} \tag{1}$$

where $S(w,f)$ and $H(w,f)$ are the frequency response of the clean speech and the room impulse response (RIR), respectively. D is the number of frames, over which the reverberation has an effect. The early reflection is due to the direct signal and some reflections that occur at earlier time. It is mostly addressed through Cepstral Mean Normalization (CMN) in the ASR system as it falls within the frame. On the other hand, the late reflection, whose effect spans over frames, can be treated as long-period noise[2][3]. Following our assumption above, we include the effects of the additive background noise $N(w,f)$ by expanding the reverberant model in Eq. (1)

$$X(w,f) \approx \quad X_E(w,f) \; + \; X_L(w,f) \; + \; N(w,f). \tag{2}$$

Enhancing the contaminated signal is defined by suppressing the effects of late reflection $X_L(w,f)$ and background noise $N(w,f)$ for ASR in noisy and reverberant conditions. Since the late reflection is treated as noise, the enhancement problem is reduced to a simple denoising problem. Thus, we can apply existing wavelet-based denoising techniques to address both the effects of late reflection and background noise based on the model in Eq. (2). In this paper, we treat the contaminants separately since the late reflection is dependent on the smearing effect of the previous D frames while the

background noise is not.

Several wavelet-based speech enhancement methods have been proposed. A typical method[4] is constructed by integrating a voice activity detection (VAD) and introducing different threshold profiles for different conditions. The use of several threshold profiles enables to cope with colored and non-stationary signals. A method which relies on the robustness of the all-pole filter in modeling the clean speech from the contaminant subspace is also proposed[5]. By clustering only the wavelet extrema, the reconstructed signal is robust to the effect of the contaminant subspace. Another method is based on filtering of the contaminated wavelet coefficients using Wiener gains[6], which we extended for dereverberation in[7]. The methods[4]-[6] are generally designed to enhance the speech waveform, but this does not guarantee an improvement in performance for the ASR application. Moreover, these methods do not address the problem of both late reflection and noise simultaneously.

In this paper, we present a method of suppressing the effects of late reflection and background noise through Wiener filtering in the wavelet domain. In the proposed scheme, prior to filtering, the wavelet parameters are optimized to improve the likelihood of the acoustic model. The optimization renders the proposed method to be more effective in the ASR application. In this paper, background noise and late reflection are jointly referred to as "contaminant signal".

The paper is organized as follows; Section II presents the proposed enhancement method based on Wiener filtering in the wavelet domain using optimized wavelet parameters. In Section III we explain the noise profile and reverberation time identification. The experimental setup and ASR evaluation results are presented in Section IV. Finally, we conclude the paper in Section V.

2. Wavelet-based Wiener Filtering

2.1 Optimizing Wavelet Parameters

A wavelet is generally expressed as

$$\Psi(\upsilon,\tau,t)=\frac{1}{\sqrt{\upsilon}}\Psi\left(\frac{t-\tau}{\upsilon}\right), \tag{3}$$

where t denotes time, υ and τ are the scaling and shifting parameters respectively. $\Psi\left(\frac{t-\tau}{\upsilon}\right)$ is often referred to as the mother wavelet. Assuming that we deal with real-valued signal, the wavelet transform (WT) is defined as

$$F(\upsilon,\tau)=\int f(t)\Psi(\upsilon,\tau,t)dt, \tag{4}$$

where $F(\upsilon,\tau)$ is the wavelet coefficient and $f(t)$ is the time-domain function. With an appropriate training algorithm, we can optimize τ and υ so that the wavelet captures specific characteristics of a certain signal of interest. The resulting wavelet is sensitive in detecting the presence of this signal given any arbitrary signal. In the wavelet filtering method, we are interested in detecting the power of clean speech, noise and late reflection given an observed contaminant. Thus, we optimize the wavelet parameters to detect these separately based on the AM likelihood as shown in Fig. 1.

2.1.1 Speech

Since we are interested in the speech subspace in general, optimizing a single wavelet to capture the general speech characteristics is sufficient. In Fig. 1, we illustrate the optimization of the wavelet for clean speech. Wavelet coefficients $S(\upsilon,\tau)$, extracted through Eq. (4), are converted back to the time domain $s_{\upsilon,\tau}$ through inverse wavelet transform (IWT). Likelihood scores are computed using the clean speech acoustic model λ_s, a Gaussian Mixture Model (GMM) of 64 components. This is a text independent model which captures the statistical information of the speech subspace. A greedy search process is iterated by adjusting υ and τ. The corresponding υ=a and τ=α that result to the highest score are selected.

2.1.2 Noise

The same procedure is applied to the case of noise, except for the creation of multiple profiles (i), representing different types of noise. $N(\upsilon,\tau)^{(i)}$ and $n^{(i)}_{\upsilon,\tau}$ are the wavelet and time domain of noise profile (i), respectively. Likelihood scores are computed using the corresponding noise model $\lambda_{n^{(i)}}$ (same model structure as that of λ_s). This model is trained using a noise database. The corresponding υ=$b^{(i)}$ and τ=$\beta^{(i)}$ that maximize the likelihood score are stored in the profile.

The noise database is originally composed of seven base noise, i.e. Car, Computer, Office, Crowd, Park, Mall and Vacuum cleaner. To generalize to a variety of noise char-

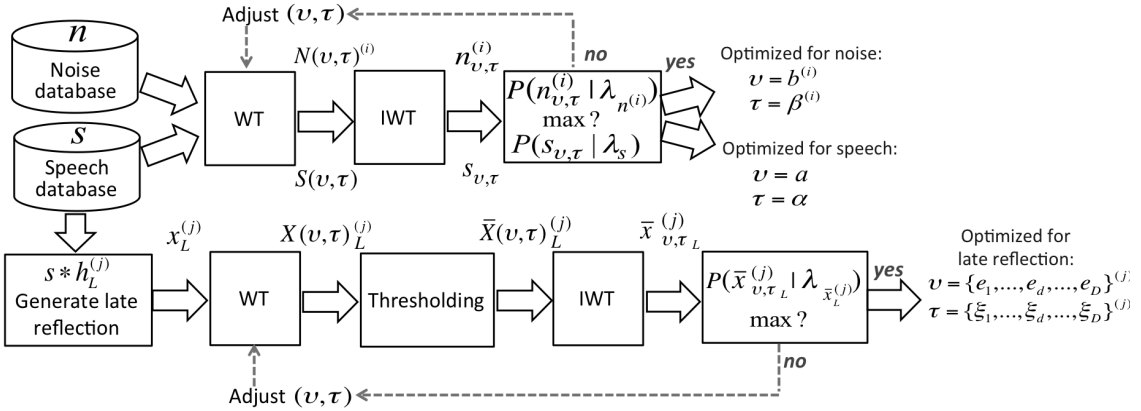

図 1 Wavelet parameter optimization scheme.

acteristics, additional entries are made by combining different types of the base noise. To remove redundancy and suppress the increase of the entries, we measure the correlation of the resulting combinations and select the ones that are less correlated with existing noise entries. Thus, the expanded noise database referred to as noise profiles will provide more degree of freedom in characterizing various noise distributions.

2.1.3 Late reflection

In the case of the late reflection in Fig. 1 (bottom), D templates for every reverberation time T_{60} (j) are to be optimized for both scale $(\upsilon_1,...\upsilon_D)^{(j)}$ and shift $(\tau_1,...,\tau_D)^{(j)}$. These correspond to D preceding frames that cause smearing to the current frame of interest. We note that the effect of smearing is not constant, thus D templates are created. By estimating the reverberation time T_{60} (j), we can generate the impulse response and its corresponding late reflection coefficients $h^{(j)}_L$[7]. Then, late reflection observations $x^{(j)}_L$ are generated by convolving the clean speech with $h^{(j)}_L$. Next, wavelet coefficients $X(\upsilon,\tau)^{(j)}_L$ are extracted through WT. In order to make $X(\upsilon,\tau)^{(j)}_L$ void of speech characteristics, thresholding is applied to $X(\upsilon,\tau)^{(j)}_L$. Speech energy is characterized with high coefficient values[8)4)] and thresholding sets these coefficients to zero,

$$\bar{X}(\upsilon,\tau)^{(j)}_L = \begin{cases} 0 & ,\mid X(\upsilon,\tau)^{(j)}_L \mid > \ thr \\ X(\upsilon,\tau)^{(j)}_L & ,\mid X(\upsilon,\tau)^{(j)}_L \mid \leq \ thr \end{cases} \tag{5}$$

thr is calculated similar to that in[8]. The thresholded signal is converted back to time domain $\bar{x}^{(j)}_{\upsilon,\tau_L}$ and evaluated against a late reflection model $\lambda_{\bar{x}^{(j)}_L}$. The parameters υ and τ are adjusted and the corresponding υ=$\{e_1,...,e_D\}^{(j)}$ and τ=$\{\xi_1,...,\xi_D\}^{(j)}$ that result to the highest likelihood score are selected. We note that $\lambda_{\bar{x}^{(j)}_L}$ is trained using the synthetically generated late reflection data (during training) with thresholding applied.

2.2 Filtering Using Wiener Gain

The general expression of the Wiener gain at window frame w and band m for background noise and late reflection are expressed as

$$\kappa^{N}_{wm} = \frac{S(\upsilon,\tau)^2_{wm}}{S(\upsilon,\tau)^2_{wm} + N(\upsilon,\tau)^2_{wm}} \tag{6}$$

and

$$\kappa^{X_L}_{wm} = \frac{S(\upsilon,\tau)^2_{wm}}{S(\upsilon,\tau)^2_{wm} + X_L(\upsilon,\tau)^2_{wm}}, \tag{7}$$

where $S(\upsilon,\tau)^2_{wm}$, $N(\upsilon,\tau)^2_{wm}$ and $X_L(\upsilon,\tau)^2_{wm}$ are wavelet power estimates for the clean

Vol.2011-SLP-87 No.14
2011/7/23

speech, noise, and late reflection, respectively. By using the optimized values for υ and τ as described in Section II-A, we can compute the respective power estimates directly from the observed contaminated signal $X(\upsilon,\tau)_{wm}$. Thus, the speech power estimate becomes

$$S(\upsilon,\tau)^2_{wm} \approx X(a,\alpha)^2_{wm}, \tag{8}$$

the noise power estimate $N(\upsilon,\tau)^2_{wm}$ as

$$N(\upsilon,\tau)^2_{wm} \approx X(b^{(i)},\beta^{(i)})^2_{wm}, \tag{9}$$

and the late reflection estimate $X_L(\upsilon,\tau)^2_{wm}$ as

$$X_L(e_d^{(j)},\xi_d^{(j)})^2_{wm} \approx \begin{cases} X(e_1^{(j)},\xi_1^{(j)})^2_{wm}, & d=1 \\ \dfrac{\sum_{k=1}^{d-1} X(e_k^{(j)},\xi_k^{(j)})^2_{wm}}{d-1} & + \\ X(e_d^{(j)},\xi_d^{(j)})^2_{wm}, & \text{otherwise,} \end{cases} \tag{10}$$

Wiener filtering is conducted by weighting the contaminated wavelet coefficient $X(\upsilon,\tau)_{wm}$ with the Wiener gain as,

$$X(\upsilon,\tau)_{wm}(enhanced) = X(\upsilon,\tau)_{wm} \ . \ \kappa_{wm}, \tag{11}$$

where we define

$$\kappa_{wm} = \frac{\kappa^N_{wm} + \kappa^{X_L}_{wm}}{2}. \tag{12}$$

Although this is not a direct calculation of the Wiener gain based on the combined effects of both noise and late reflection, we used Eq. (12) for reason of tractability. In Eq. (11), the Wiener weight κ_{wm} dictates the degree of suppression of the contaminant to the observed signal at particular frame w and band m. If the contaminant power estimate is greater than the estimate of the speech power, then κ_{wm} for that band may be set to zero or a small value. This attenuates the effect of contamination. On the other hand, if the power of the clean speech estimate is greater, the Wiener gain will emphasize its effect. The enhanced wavelet coefficients are converted back to the time domain through IWT and given to the ASR process.

3. Noise Profile and T_{60} Identification

Each noise profile (i) and reverberation time T_{60} (j) has corresponding optimized wavelet parameters ($b^{(i)}$, $\beta^{(i)}$), $\{e_1,...,e_D\}^{(j)}$ and $\{\xi_1,...,\xi_D\}^{(j)}$ as shown in Section II-A. For actual ASR, it is necessary to identify the profile that corrupts the speech signal to retrieve the appropriate parameters. To identify the noise profile (i), a GMM-based classifier is employed. The GMMs ($\lambda_{n^{(i)}}$) are same as used in optimizing the wavelet parameters for the noise profiles discussed in Section II-A. Prior to ASR, high-energy frames are removed from the input noisy speech and the remaining noise segments are evaluated with the GMMs. Subsequently, the profile (i) that leads to the best likelihood is selected. The same procedure is applied to the identification of T_{60} (j), using the GMM classifier $\lambda^{(j)}_{\bar{x}_L}$ trained with the synthetically generated late reflection data. We have found out that the identification works well even with only a few frames of data.

4. Experimental Evaluations

We have evaluated the proposed method in large vocabulary continuous speech recognition (LVCSR). The training database is the Japanese Newspaper Article Sentence (JNAS) corpus with a total of approximately 60 hours of speech. The test set is composed of 200 sentences uttered by 50 speakers. The vocabulary size is 20K and the language model is a standard word trigram model.

Speech is processed using 25ms-frame with 10ms. shift. The features used are 12-order MFCCs, ΔMFCCs, and ΔPower. The AM is a phonetically tied mixture (PTM) HMMs with 8256 Gaussians in total. It is trained using the speech database with superimposition of Gaussian noise, that is different from those in the noise profiles[9)10)]. We note that in our proposed method, we use only a single AM in ASR for different noise and SNR conditions. We used seven types of real noise (base noise) in the NAIST database[10)]: Car, Computer, Office, Crowd, Park, Mall and Vacuum cleaner. As the result of combination of the base noise entries, 20 noise profiles are generated. We considered reverberation time T_{60} from 100ms. to 500ms. with 100ms. interval. In the experiments, we compare the proposed method against modified wavelet-based methods[4)-6)] in dealing with the reverberation problem[7)]. Then we perform post-processing

Vol.2011-SLP-87 No.14
2011/7/23

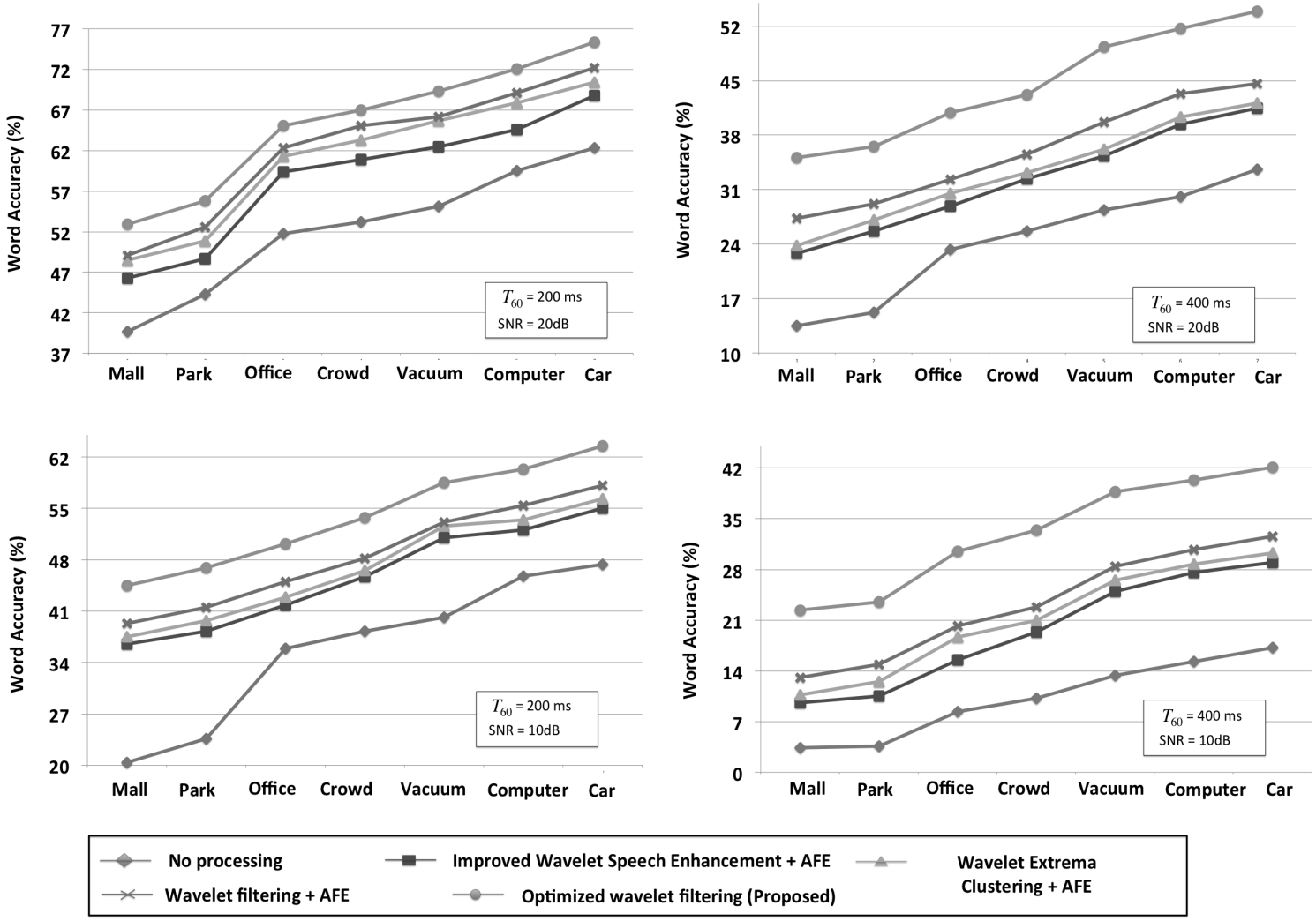

図 2 Recognition Performance.

using the ETSI advanced front-end (AFE)[11] to deal with the background noise for these methods.

In Fig. 2, we show the ASR performance in word accuracy for different noise types, SNRs (10, 20dB) and reverberation time (200, 400ms.). We note that when a particular noise-type is being evaluated, it is held-out during noise profile generation. (A) is the result when the contaminated data is not processed and recognized using an AM re-trained with the same condition. (B) is the result when processed with the improved wavelet-based enhancement that incorporates VAD and threshold profiles[4]. Another method based on extrema clustering[5] is evaluated in (C). The result of wavelet filtering without optimization[6][7] is shown in (D), while the result of the proposed method which incorporates both late reflection and background noise is given in (E). The results in Fig. 2 show that the proposed method outperforms existing wavelet-based methods in all cases[4]-[7]. By optimizing the wavelet parameters, the enhancement process is tuned to improving the acoustic model likelihood. As a result, the proposed method becomes more effective in the ASR application.

5. Conclusion

The proposed wavelet-based Wiener filtering optimizes the wavelet parameters to effectively estimate the power of the clean speech, noise, and late reflection. This optimization is based on the AM likelihood, and results to a more accurate Wiener gain estimate in suppressing the contaminant signal. Currently, we deal with simple additive background noise. In the future, we will further investigate its convolutive effect. This scenario occurs when the noise source is located at a considerable distance from the microphone.

参 考 文 献

1) R. Gomez and T. Kawahara, "Robust Speech Recognition Based on Dereverberation Parameter Optimization Using Acoustic Model Likelihood" *IEEE Trans. on Audio, Speech and Lang. Proc.*, Sept. 2010
2) R. Gomez et.al. , "Distant-talking Robust Speech Recognition Using Late Reflection Components of Room Impulse Response" *ICASSP*, 2008.
3) R. Gomez and T. Kawahara, "Optimization of Dereverberation Parameters based on Likelihood of Speech Recognizer" *Interspeech*, 2009.
4) H. Sheikhzadeh and H. Abutalebi, "An Improved Wavelet-based Speech Enhancement System" *In Proceedings of Eurospeech*, 2001.
5) S. Griebel and M. Brandstein, "Wavelet Transform Extrema Clustering for Multichannel Speech Dereverberation" *IEEE Workshop on Acoustic Echo and Noise Control*, 1999
6) E. Ambikairajah et. al., "Wavelet Transform-based Speech Enhancement" *In Proceedings of ICSLP*, 1998.
7) R. Gomez, T. Kawahara, "Optimizing Wavelet Parameters for Dereverberation in Automatic Speech Recognition" *In Proceedings of APSIPA*, 2010.
8) D.L. Donoho, "Denoising by soft thresholding", *IEEE Trans. Info. Theory* 1995.
9) D.V. Compernolle, "Noise Adaptation in a Hidden Markov Model Speech Recognition System" *Computer Speech and Language* 1989.
10) S. Yamade, K. Matsunami, A. Baba, A. Lee, H. Saruwatari and K. Shikano, "Spectral subtraction in noisy environments applied to speaker adaptation based on HMM Sufficient Statistics", *In Proceedings of ICSLP*, 2000.
11) Advanced Front-End Feature Extraction Algorithm, *ETSI Standard Document ES 202 050*, 2002.