Speech Event Detection using Multiband Modulation Energy

Georgios Evangelopoulos and Petros Maragos

National Technical University of Athens, School of E.C.E, Athens 15773, Greece.

[gevag,maragos]@cs.ntua.gr

Abstract

The need for efficient, sophisticated features for speech event detection is inherent in state of the art processing, enhancement and recognition systems. We explore ideas and techniques from non-linear speech modeling and analysis, like modulations and multiband filtering and propose new energy and spectral content features derived through filtering in multiple frequency bands and tracking dominant modulation energy in terms of the Teager-Kaiser Energy of separate AM-FM components. We present a detection-theoretic motivation and incorporate them in two detection schemes namely word boundary and voice activity detection. The modulation approach demonstrated noisy speech endpoint detection accuracy, reaching $\sim$40% error reduction on NTIMIT. In a voice activity scheme, improvement in overall misclassification error of a high hit-rate detector reached 7.5% on Aurora 2 and 9.5% on Aurora 3 databases.

1. Introduction

Detecting speech presence can be thought of either as a direct problem of event labeling in silence and noise or as an indirect one of voice activity detection. Separation of speech from background noise is a special case of the general problems of speech segmentation and event detection and is important for speech recognition, coding, processing and transmission. Critical processing reduction is achieved by selecting the useful speech segments of a recorded signal, while recognition systems front-ends require highly accurate formation of speech patterns. Speech detection is incorporated in labeling large databases and in various enhancement and manipulation techniques like noise spectrum estimation [1] and frame dropping [2], noise reduction, echo cancelation, energy normalization and silence compression. In telecoms it is applied for selective, real-time speech transmission over networks.

Modern detection approaches focus either on the development of sophisticated features or on the decision logic complexity. Novel features for noisy speech labeling are inspired by exploring alternative signal properties. Apart from energy and zero-crossings rate, literature includes 'periodicity' and jitter, pitch stability, spatial signal correlation, spectral entropy, cepstral features, LPC residual, alternative energy measures [3], temporal power envelope [1], spectral divergence [2, 4], and time-frequency parameters through multiband analysis [5]. Recently the statistical framework gains interest with properties of the speech statistics being used along with optimized likelihood ratio rules [6].

Our approach involves new time-domain signal representations derived using multiple frequency band demodulation [7] of the signal in AM-FM components through energy separation [8]. A maximum average energy tracking process over the various frequency bands is used to yield short-time features of multiband signal modulation energy and demodulated instant amplitude and frequency. To verify the effectiveness of the modulation-based features for speech event detection, we incorporated them, in an endpoint locating threshold-based algorithm and a voice activity algorithm based on adaptive optimum thresholds for noisy speech detection [2]. These features consistently improved detection performance under various noise levels and conditions on large databases.

2. Energy Operators and Multiband Modulations

Indications of multi-scale modulations during speech production, led to the proposal of the AM-FM modulation model for speech [8]. Demodulation of a real-valued AM-FM signal $x(t) = a(t)\cos\left(\int_0^t \omega(\tau)d\tau\right)$ with time varying amplitude envelope $a(t)$ and instantaneous frequency $\omega(t)$ signals, can be approached using the non-linear Teager -Kaiser differential energy operator. For continuous-time signals $x(t)$, this operator is $\Psi[x(t)] \equiv [\dot{x}(t)]^2 - x(t)\ddot{x}(t)$, where $\dot{x}(t) = dx(t)/dt$. The energy operator Ψ can track the instantaneous energy of a source producing an oscillation. Applied to an AM-FM signal, Ψ yields with negligible approximation error the instantaneous source energy, i.e. $\Psi[x(t)] \approx a^2(t)\omega^2(t)$. An efficient AM-FM demodulation scheme based on Ψ, the *energy separation algorithm* (ESA) [8], separates the instantaneous energy into its amplitude and frequency components. The algorithm's discrete counterpart is simple, computationally efficient and has an excellent (almost instantaneous) time resolution.

In order to apply demodulation through ESA in a AM–FM modeled speech signal, it is necessary to filter the signal and isolate specific frequency bands. The *multiband demodulation analysis* (MDA) scheme was introduced in [7] as a way of capturing modulations in the presence of noise.

2.1. Detection-theoretic motivations and optimality

Effective labeling of speech activity must take into account both the energy level of the excitation and its frequency content. Teager's definition of the energy of a signal as the energy produced by its generating source fits that framework by simultaneously counting spectral and magnitude level.

Consider the sum of modulated sines speech model $s[n] = \sum_{k=1}^{K} A_k[n]\cos(\Omega_{ck} \cdot n + \Phi_k[n])$ where k is the resonance index and K the number of speech formants. Suppose now that a single AM-FM signal is present, i.e. $K = 1$, by capturing a single modulation with a sufficiently narrowband Gabor filter. The carrier frequency Ω_{ck} may be assumed known and approximated via the central frequency of the Gabor filter. Formulating the problem for simplicity as the detection of a sinusoid with unknown, non-random parameters and frame-wise stationary amplitude and phase in Gaussian noise of unknown variance, the two hypothesis are: $H_0 : X[n] = W[n], H_1 : X[n] =$

$W[n] + A\cos(\Omega_c n + \Phi) + B$, for each frame of length N.

By using the *Maximum Likelihood Estimates* $[\hat{A}, \hat{\Phi}, \hat{B}, \hat{\sigma}_1^2, \hat{\sigma}_0^2]$ for the unknown parameters, it is straightforward to prove that for the Gaussian distribution $p(X|H_1)$ conditioned on H_1:

$$\ln p(X|H_1) \approx N\frac{\hat{A}^2}{4\hat{\sigma}_1^2} + \frac{N}{2\hat{\sigma}_1^2}(\hat{B}^2 - \hat{\sigma}_0^2) - \frac{N}{2}\ln 2\pi\hat{\sigma}_1^2 \quad (1)$$

We decide in favor of hypothesis H_i that maximizes the log-likelihood function and to account for the different number of estimated parameters the *Minimum Description Length* criterion. In detail we choose H_i that maximizes $\mathrm{MDL}(i) = \ln p(X|H_i) - \frac{n_i}{2}\ln N,\ i = 1, 2$ with $n_i = [1, 4]$ the parameter vector dimensionality.

Estimation of Ω_c using a Gaussian (or rectangular) window of frequency spread σ_g is governed by certain mean duration-average frequency uncertainty relations which after a few manipulations sum to $\Omega_c^2 + \sigma_g^2 \geqslant 1/4N^2$. By using the lower uncertainty bound with $n_1 = 4$ for the $\mathrm{MDL}(1)$ we have $-\ln N^{\frac{n_i}{2}} = \ln 4 + \ln(\Omega_c^2 + \sigma_g^2)$ and we can then construct a rule for the test on the sinusoid-speech component detection:

$$N\frac{\hat{A}^2}{4\hat{\sigma}_1^2} + \ln(\Omega_c^2 + \sigma_g^2) + \underset{H_0}{\overset{H_1}{\gtrless}} \mathcal{O}(\hat{B}, \hat{\sigma}_1^2, \hat{\sigma}_0^2, N) \quad (2)$$

where $\mathcal{O}$ a function of statistics on the analysis frame and $\hat{A}^2/2\hat{\sigma}_1^2$ is the SNR. The above rational also applies for detecting one out of K sinusoids with different carrier frequencies of K Gabor filters. We then have to test $K + 1$ hypotheses by maximizing the MDL criterion and labeling a frame as noise if $\mathrm{MDL}(0) > \mathrm{MDL(i)}, \forall i \neq 0$.

From [7] the expected value of the energy operator on a filtered AM-FM signal-plus-noise can be approximated by $\Psi(X[n]) \approx A[n]^2|H(\Omega_c)|^2[(\Omega_c + \partial\Phi[n]/\partial n)^2 + \Gamma_c]$, where Ω_c and H are respectively the filter's central frequency and frequency response and Γ_c is a constant standing for the averaged filtered noise power. In our case this approximation yields $\Psi(X[n]) \approx A^2(\Omega_c^2 + \Gamma_c)\cdot|H(\Omega_c)|^2$ and by taking logarithms:

$$\ln\Psi(X[n]) \approx \ln A^2 + \ln(\Omega_c^2 + \Gamma_c) + \mathrm{const.} \quad (3)$$

Comparing Eq. (2) and (3) we notice the amplitude-frequency product components and the constants depending on the average bandpass noise inside the logarithms. These similarities despite the approximations on the problem, give an insight on the role of the energy operator and the ESA estimates on a channel decision-speech detection process.

2.2. Modulation-based features

Modulation bands are obtained through a linearly-spaced bank of K Gabor bandpass filters and the discrete energy operator Ψ_d [8] is applied as nonlinear energy measurement. For each signal frame, m, short-time representations of the dominant modulation components are obtained by tracking, in the multi-dimensional feature space consisting of the filter responses on the signal s, the *maximum average Teager Energy* (MTE):

$$\mathrm{MTE}(m) = \max_{1\leq k\leq K}\frac{1}{N}\sum_{n=1}^{N}\Psi_d[(s * h_k)(n)] \quad (4)$$

where $*$ denotes convolution, n is the sample index with $(m-1)N + 1 \leq n \leq mN$ and h_k the impulse response of the k_{th} filter. The filter $j = \arg\max_k(\mathrm{MTE})$ is submitted to demodulation via ESA to derive the mean Instant Amplitude (MIA) and mean Instant Frequency (MIF) features. MTE may be thought of as the dominant signal *modulation energy*, capturing the joint amplitude-frequency information inherent in speech activity.

Classic		Multiband	Modulation	
mAA,ZR	mSA,ZR	MAA,MZR	MTE,MIF	MTE
56.1	66.6	51.5	73.5	73.1

Table 1: *Detected speech endpoints (%) on NTIMIT*

3. Endpoint Detection in Noise

Modulation-based features were used instead of conventional mean absolute amplitude (mAA) and zero-crossings rate (ZR) in a classic double threshold endpoint detection scheme [9]. From the first 100 ms, which are a-priori assumed silence, the mean μ_{sif} and standard deviation σ_{sif} values of the 'silent' MIF are computed along with the maximum MTE values for silence, $S_{\max}$, and for the whole signal, $P_{\max}$. Threshold rules are constructed using κ, λ as weighting constants, according to:

$$\gamma_{\mathrm{f}} = \mu_{\mathrm{sif}} + \kappa\sigma_{\mathrm{sif}},\ \gamma_{\mathrm{d}} = \min(T_1, T_2),\ \gamma_{\mathrm{u}} = 5\cdot\gamma_{\mathrm{d}} \quad (5)$$

$$T_1 = \lambda P_{\max} + (1-\lambda)S_{\max} \quad , \quad T_2 = 3\cdot S_{\max} \quad (6)$$

The double energy check, searching for the boundary points where a high threshold γ_{u} is exceeded after a low-stricter one γ_{d}, detects the main, usually voiced duration of the speech signal. These initial endpoints are refined to include strong spectral unvoiced activity using the γ_{f} frequency threshold. In our tests we set $\lambda = 0.02, \kappa = 1$.

We tested features for endpoint detection under real noise conditions on the NTIMIT database. The task was detection of phrase boundaries, ignoring in-between activity, compared to manually labeled boundaries. An error in detection was considered if the boundaries were misplaced for more than 60 ms. Detection percentages for the whole set, using various feature approaches can be seen in Table 3. mSA refers to mean square amplitude while MAA, MZR are multiband versions of conventional amplitude (max average amplitude) and zero-crossings rate (rate at the band with max mean filtered envelope). Decrease of the average detection error, compared to the classic features, was 38.7% with the use of the MTE feature and 40.1% after refinement with MIF.

To evaluate results independent from the empirically defined error thresholds we used a simple convention to produce curves that approximate ROC curves, which we will call AROC (*Approximate ROC*). We set a tight threshold at 30 ms for lost-phoneme tolerance and let the boundary misplacement threshold vary from 2 to 150 ms. These thresholds, yield a measure related to the Probability of False Alarm (PF). The two quantities are not equal but they are connected by a one-two-one, monotonically increasing unknown function as increase in the threshold increases the PF by *some* amount. In Fig. 1 the AROC curves are plotted for the modulation, classic and the multiband-classic features.

4. Voice Activity Detection

Any VAD system classifies incoming signal frames based on feature estimation in two classes: speech and non-speech events (pauses, silence, background noise). The recently developed

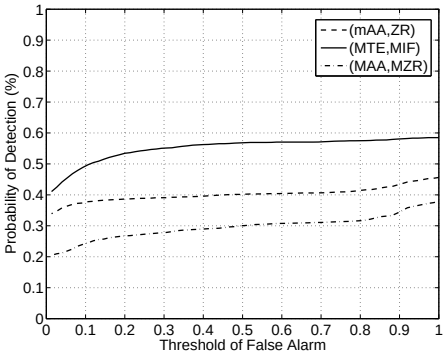


Figure 1: *Approximate ROC curves on NTIMIT. The threshold of False Alarm is a relative time interval of tolerance in endpoint estimation.*

and highly accurate VAD system in [2, 4] implicates the use a feature termed *Long Term Spectral Divergence* (LTSD) and is based on adaptive thresholds and noise parameter updating. The feature quantifies speech divergence from background noise and is in essence an energy measure that retains spectral information of strong spectral components over neighboring frames. The LTSD VAD was extensively and successfully tested under varying noise conditions against standard VAD [10], using common evaluation methods and recognition accuracy.

4.1. Modulation Energy Detection

To evaluate the MTE as a speech event detector in a VAD system, we adapted the LTSD-based algorithm changing the core feature with the proposed modulation-based MTE in two alternative expressions. The signal is frame-processed and during a short initialization period the initial noise characteristics are learned. After feature computation, the level difference in dB from the respective background noise feature is compared to an adaptive threshold $\gamma \in [\gamma_0, \gamma_1]$:

$$\gamma = \gamma_0 + (\gamma_1 - \gamma_0)(E - E_0)/(E_1 - E_0) \qquad (7)$$

where E the background noise energy and the threshold interval boundaries depend on the cleanest E_0 and noisiest E_1 energies, computed during the initialization period from the considered database. The noise feature is initialized and adapted whenever silence or pause is detected, by averaging in a small frame neighborhood.

To measure modulation 'divergences' in the spirit of the LTSD, we use features based on the MTE: 1) *Multiband Teager Energy Divergence* (MTED) The multiband max average Teager Energy MTE as previously used, compared to the respective MTEW for the background noise:

$$\text{MTED}(m) = 10\log_{10}\left(\text{MTE}(m)/\text{MTEW}\right) \qquad (8)$$

The MTED is measuring the divergence between the multiband MTE of a frame and the corresponding noise feature. This is conceptually the same as the endpoint detection algorithm of Sec.3, comparing MTE level difference. 2) *Long-term Multiband Teager Energy Divergence* (LTED), where the MTE is locally maximized in a neighborhood of $2L$ frames resulting in a dilated and normalized, with respect to the background noise, version:

$$\text{LTED}(m) = 10\log_{10}\left(\max_{l}\{\text{MTE}(m+l)\}/\text{MTEW}\right) \qquad (9)$$

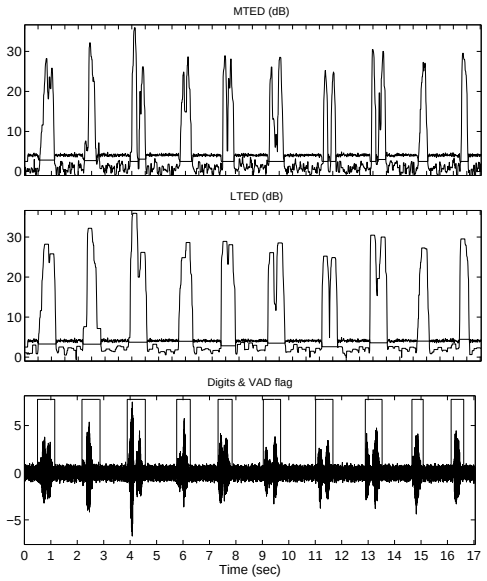


Figure 2: *Examples of features based on maximum average Teager Energy (MTE) for VAD. The VAD flags were almost identical using both modulation features, LTED and MTED.*

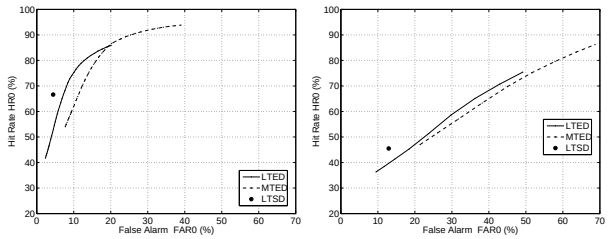
with $-L \leq l \leq L$ defining the order of dependence. In Fig. 2 we present an example of the proposed features for VAD on an example by the Aurora 3 database (quiet, hands-free mic., digit sequence), for frames of 25ms with 10ms shifts and a bank of 25 Gabor filters. Superimposed is the adaptive threshold signalling voice activity.

4.2. Experimental evaluation

The experimental framework involves comparing detection performance of the LTSD and the MTE-based VADs, under varying noise conditions on the Aurora 2 [11] (70070 utterances) and Aurora 3 (4914 utterances) databases. Evaluation is based on classification errors at different SNRs [1, 2, 10] using some reference labeling of the clean digit dataset. In our experiments automatic speech recognition was used to segment and label speech and silence events on the databases. High recognition rate results on the clean sequences defined the ground truth. Briefly, for the Aurora 2 set, the training was done using 32 mixtures, 18 states and the 39-long feature vector $[\text{MFCC}, \log \text{E}, \Delta, \Delta\Delta]^{\text{T}}$ on the clean-train scenario with the test run on the clean data achieving a 99.6% word accuracy. For Aurora 3, training was done with 16 mixtures, 16 states and the same features. The 1522 subset for the well matched test scenario was used with a 93.7% recognition accuracy.

For the reference LTSD-based VAD we used the specifications reported in [2], while for the proposed VADs we determined the optimum thresholds by means of the ROC curves. At Fig. 3 we present these curves for the noisiest and cleanest sets for the MTED and LTED-based VADs. We chose the thresholds that correspond to the points of the curves with minimum distance from the upper left, ideal working point, corner. This led to $\gamma_0 = 24$ dB, $\gamma_1 = 0.5$ dB for the MTED VAD and $\gamma_0 = 32$ dB, $\gamma_1 = 2$ dB for the LTED on the Aurora 2 set. The tests on Aurora 3 were conducted with the same pair of thresholds ($\gamma_0 = 6$ dB, $\gamma_1 = 2.5$ dB) for all three features.

Performance of the VADs was evaluated with respect to the speech Hit Rate HR1, defined as the ratio of the detected



(a) (b)

Figure 3: *ROC curves for speech detection performance in (a) clean and (b) noisiest (-5 dB) Aurora 2 sets for the MTE based VADs. The operating point of the LTSD-VAD is also depicted.*

speech frames to the total number of speech frames and the non-speech Hit Rate HR0, defined analogously for silence. Complementary to these quantities the False Alarm Rates, FAR1 and FAR0 of the decision for speech or non-speech are defined. The rates HR1 and HR0, are considered of equal importance due to misclassification errors taking place in both in speech and non-speech periods. This may be quantified by the L_2 norm of false alarms. We aim to minimize the *overall false alarm error norm*:

$$\|(\text{FAR0,FAR1})\| = \left[(1-\text{HR0})^2 + (1-\text{HR1})^2\right]^{1/2} \quad (10)$$

Statistically this measure expresses the average performance of the detector as $\sqrt{2}$ times the rms norm of the false alarms, while geometrically it is the shortest Euclidian distance from the ideal operating point (upper left corner) on the ROC plot of a detector (HR=100, FAR=0) (see also Fig. 3). In Fig. 4 the error norm is presented for the two datasets and the three VADs as a function of decreasing SNR.

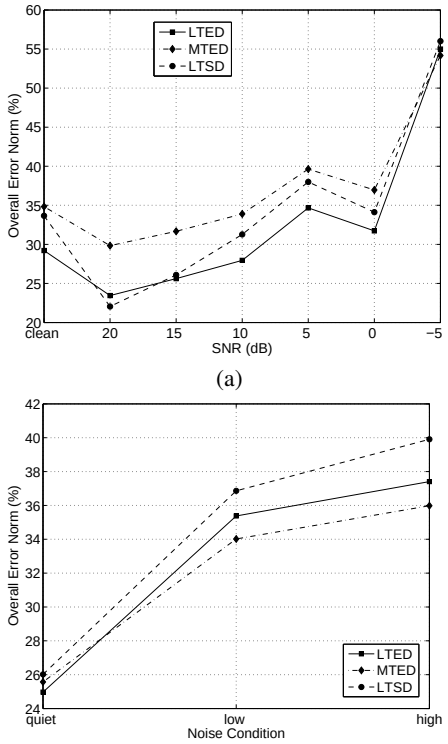
On the Aurora 2 tests, where the thresholds were optimally set, the MTE-based algorithms equally weigh both rates giving average hit rates above 70% on both speech and silence periods. The LTED achieved the minimum false alarm error norm, with a 7.6% decrease of the overall error over the LTSD-based VAD. In Fig.4(a) the LTED detector minimizes the error, except on 20 dB SNR, where all three features follow analogous degradations in performance under increasing SNR. On the Aurora 3 set where the detection thresholds were the same, the LTED achieves higher individual hit rate performance than LTSD and an overall decrease in error of 7.7%. Minimum false alarm error is given by the MTED feature with a relative decrease of 9.5%, while both modulation-feature based algorithms outperform the LTSD in terms of the overall error under all three noise conditions, as can be seen in Fig. 4(b). Note that the LTED feature is consistently best on both test sets.

5. Conclusions

We approached the detection of voiced events as the detection of speech modulations, tracking through multiple bands their dominant structures and measuring slowly-varying amplitude and frequency information. Modulation features were systematically verified to improve speech boundary detection and decrease average error of a robust and accurate VAD using various benchmarks. Extending these ideas to general event detection and analysis can be done through a generalized modulation description of various acoustic signals.

6. Acknowledgements

This work has been supported partially by the European NoE 'MUSCLE', the European research program 'HIWIRE' and the NTUA research program 'Protagoras'.



(a)

(b)

Figure 4: *Overall false alarm errors for voice event detection under various SNR on (a) Aurora 2 and (b) Aurora 3 (Dashed is the reference LTSD, solid and dash-dotted the MTE features).*

7. References

[1] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Trans. Speech Audio Processing*, vol. 10, no. 2, pp. 109–118, Feb. 2002.

[2] J. Ramirez, J. Segura, C. Benitez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, no. 3-4, pp. 271–287, Apr. 2004.

[3] G. Ying, C. Mitchell, and L. Jamieson, "Endpoint detection of isolated utterances based on a modified Teager energy measurement," in *Proc. IEEE ICASSP '93*, Minneapolis, MN, Apr. 1993, pp. 732–735.

[4] J. Ramirez, J. Segura, C. Benitez, A. de la Torre, and A. Rubio, "A new Kullback-Leibler VAD for speech recognition in noise," *IEEE Signal Processing Lett.*, vol. 11, no. 2, pp. 266–269, Feb. 2004.

[5] G. Wu and C. Lin, "Word boundary detection with mel-scale frequency bank in noisy environment," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 5, pp. 541–553, Sept. 2000.

[6] J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[7] A. Bovik, P. Maragos, and T. Quatieri, "AM-FM energy detection and separation in noise using multiband energy operators," *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3245–3265, Dec. 1993.

[8] P. Maragos, J. Kaiser, and T. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. Signal Processing*, vol. 41, no. 10, pp. 3024–3051, Oct. 1993.

[9] L. Rabiner and M. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Systems Tech. J.*, vol. 54, no. 2, pp. 297–315, Feb. 1975.

[10] F. Beritelli, S. Casale, G. Ruggeri, and S. Serrano, "Performance evaluation and comparison of G.729/AMR/Fuzzy voice activity detectors," *IEEE Signal Processing Lett.*, vol. 9, no. 3, pp. 85–88, Mar. 2002.

[11] H. Hirsh and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noise conditions," in *ISCA ITRW ASR2000: "ASR: Challenges for the Next Millennium"*, Paris, France, Sept. 2000.