COMPARING SPECTRUM ESTIMATORS IN SPEAKER VERIFICATION UNDER ADDITIVE NOISE DEGRADATION

C. Hanilçi[1,2], T. Kinnunen[2], R. Saeidi[2], J. Pohjalainen[3], P. Alku[3], F. Ertaş[1], J. Sandberg[4], M. Hansson-Sandsten[4]

[1] Department of Electronic Engineering Uludağ University, Bursa, Turkey
[2] School of Computing, University of Eastern Finland, Joensuu, Finland
[3] Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
[4] Mathematical Statistics, Centre for Mathematical Sciences, Lund University, Lund, Sweden
chanilci@uludag.edu.tr, tkinnu@cs.joensuu.fi, jpohjala@acoustics.hut.fi, sandberg@maths.lth.se

ABSTRACT

Different short-term spectrum estimators for speaker verification under additive noise are considered. Conventionally, mel-frequency cepstral coefficients (MFCCs) are computed from discrete Fourier transform (DFT) spectra of windowed speech frames. Recently, linear prediction (LP) and its temporally weighted variants have been substituted as the spectrum analysis method in speech and speaker recognition. In this paper, 12 different short-term spectrum estimation methods are compared for speaker verification under additive noise contamination. Experimental results conducted on NIST 2002 SRE show that the spectrum estimation method has a large effect on recognition performance and stabilized weighted LP (SWLP) and minimum variance distortionless response (MVDR) methods yield approximately 7 % and 8 % relative improvements over the standard DFT method at -10 dB SNR level of factory and babble noises, respectively in terms of equal error rate (EER).

Index Terms— spectrum estimation, speaker verification

1. INTRODUCTION

Short-term spectrum estimation is an integral part in speech and audio applications. Discrete Fourier transform (DFT) and linear prediction (LP) are the two most commonly used methods for estimating the short-term spectrum, which is subsequently transformed into a feature vector [1]. Typically, mel-frequency cepstral coefficients (MFCCs) and linear predictive cepstral coefficients (LPCCs) are used as features in speech and speaker recognition [2]. The features are used for modeling speakers or phonemic information using, e.g., Gaussian mixture models [3, 4, 5].

Two major challenges in speaker recognition are to deal with channel mismatch and additive noise contamination. Channel mismatch occurs when the training and test handsets or channels are different (e.g., landline versus wireless). In additive noise contamination, recognition accuracy decreases because other environmental sounds get added to the speech signal. A number of techniques have been proposed for compensating these adverse effects. Speech enhancement techniques such as spectral subtraction [6] can be applied prior to feature extraction. Feature domain methods such as RASTA filtering [7] and cepstral mean and variance normalization (CMVN), in turn, improve robustness against channel mismatch or additive noise. Score normalization [8] is commonly used for dealing with score variabilities across different conditions or speakers. These methods are usually combined in a full speaker recognition system.

Recently, conventional DFT spectrum estimation was compared to LP-based methods in speaker verification under additive noise [9]. It was reported that LP-based spectrum estimators outperform the DFT method in terms of recognition accuracy. In another recent study, a non-parametric multitaper spectrum estimator was used for MFCC extraction in speaker recognition [10, 11]. In that study, the multitaper method also outperformed the standard DFT method. In [12], the minimum variance distortionless response (MVDR) method [13] was proposed for MFCC extraction in automatic speech recognition (ASR) with promising results. In [14], the MVDR method was applied to speaker verification. It was reported that baseline MFCCs outperforms the proposed method whereas fusion of two systems improves the recognition accuracy.

Another LP-based method, regularized LP (RLP), was proposed in [15] to improve spectral envelope estimation by penalizing rapid spectral changes in the conventional LP method. Another simple technique for spectrum envelope estimation method uses iterative cepstral smoothing (ICS) to remove harmonic information from a DFT spectrum [16]. To the best of our knowledge, the RLP and ICS methods have not been previously applied to speaker recognition. The LP variants and multitaper methods in [9, 10, 11] were compared with different speaker recognition set-ups.

In this paper, we compare a wide range of different short-term spectrum estimation methods for MFCC feature extraction on speaker recognition performance under additive noise contamination. Seven different all-pole spectrum estimation methods, the multi-taper method with three different window functions and the ICS based spectrum estimation method are evaluated in comparison to a standard DFT method. The all-pole methods are conventional LP, weighted linear prediction (WLP), stabilized WLP (SWLP) [17], eXtended weighted linear prediction (XLP) and its stabilized version (SXLP) [18], MVDR and RLP.

2. SPECTRUM ESTIMATION METHODS

2.1. Nonparametric spectrum estimators

In the conventional **DFT** spectrum estimator [1], the power spectrum of windowed speech frame is computed as:

$$S_{\mathrm{DFT}}(f) = \left| \sum_{n=0}^{N-1} w(n)x(n)e^{-j2\pi nf/N} \right|^2, \qquad (1)$$

The work of T. Kinnunen and J. Pohjalainen was supported by the Academy of Finland (project no. 132129 and 127345)

where f is the frequency, $w(n)$ is the window function (here, Hamming) and $x(n)$ is a speech frame of N samples. Windowing reduces the bias but variance remains high [10]. To reduce variance, the *multi-taper* method can be used instead [10, 11]:

$$S_{\mathrm{MT}}(f) = \sum_{k=1}^{K} \lambda_k \left| \sum_{n=0}^{N-1} w_k(n) x(n) e^{-j2\pi nf/N} \right|^2 \quad (2)$$

Here, K is the number of tapers and $w_k(n),\ k = 1,\ \ldots,\ K$ are the tapers with weights λ_k. Thus, the multitaper spectrum estimate is a weighted average of K individual spectra. In the literature, there exists a number of different tapers for spectrum estimation. In this study we consider the *Thomson*, *multipeak* and *sine weighted cepstrum estimator* (SWCE) tapers as in [11].

The **ICS** method [16] is based on the cepstral smoothing technique. First, the DFT spectrum of the analysis frame $S(f)$ is computed using (1). The spectral envelope at iteration i, $A_i(f)$ is then updated as the maximum of the original spectrum and the current spectral envelope, $C_{i-1}(f)$,

$$A_i(f) = \max(\log|S(f)|, C_{i-1}(f)), \quad (3)$$

where $C_i(f)$ is the cepstrally smoothed spectrum at the i^{th} iteration. $A_0(f) = \log|S(f)|$ is used to compute $C_0(f)$ for the initial setting as the starting point.

2.2. Parametric all-pole spectrum estimators

A p^{th} order **LP** analysis [1] assumes that each speech sample at a given discrete time index n, can be estimated as a linear combination of its p previous samples, $\hat{x}(n) = \sum_{k=1}^{p} a_k x(n-k)$, where $x(n)$ is the original speech sample and $\hat{x}(n)$ is the predicted sample. The objective of LP analysis is to find the predictor coefficients, a_k, by minimizing the energy of the prediction residual, $E = \sum_n (x(n) - \sum_{k=1}^{p} a_k x(n-k))^2$. Given p prediction coefficients, $a_k,\ k = 1,\ 2,\ \ldots,\ p$, the LP spectral envelope is computed by:

$$S_{\mathrm{LP}}(f) = \frac{1}{|1 - \sum_{k=1}^{p} a_k e^{-j2\pi fk}|^2}. \quad (4)$$

In **WLP** [17], predictor coefficients, b_k, are computed by minimizing the energy of a weighted squared error signal, $E = \sum_n (x(n) - \sum_{k=1}^{p} b_k x(n-k))^2 W_n$, where W_n is the short-time energy (STE) of the signal history, $W_n = \sum_{i=1}^{M} x^2(n-i)$ and M is the length of the STE window. The WLP method corresponds to conventional LP analysis for the case of weighting function chosen as $W_n = d$, for all n and $d \neq 0$.

Conventional autocorrelation LP guarantees the stability of the all-pole model (i.e. filter poles are inside the unit circle). Filter stability is essential in speech coding and synthesis applications. However, such a guarantee does not exist for the WLP method. Thus, stabilized WLP (**SWLP**) was proposed in [17] which uses a recursive weighting function.

In **XLP** [18], the predictor coefficients, c_k, are computed by minimizing the following objective function:

$$E_{\mathrm{XLP}} = \sum_n (x(n) Z_{n,0} - \sum_{k=1}^{p} c_k x(n-k) Z_{n,k})^2, \quad (5)$$

where $Z_{n,j} = \frac{m-1}{m} Z_{n-1,j} + \frac{1}{m}(|x(n)| + |x(n-j)|)$ and $Z_{n,j} = 0$ for $j < 0$. The stabilized version of XLP, **SXLP**, corresponds to the case of weighting function $Z_{n,j}$ chosen as $Z'_{n,j} = \max(Z_{n,j}, Z_{n-1,j-1})$.

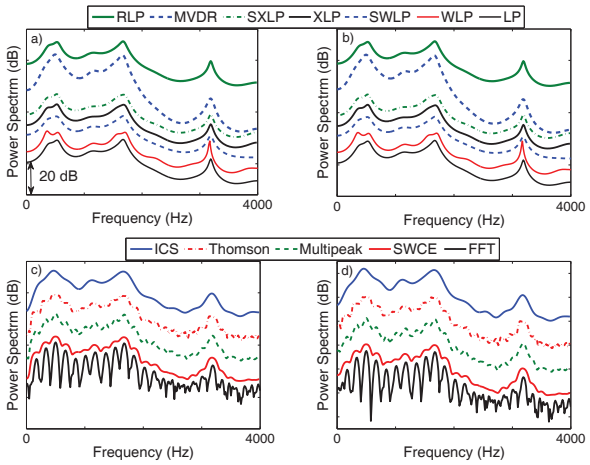

Fig. 1. Short-term spectra of an original speech frame taken from the NIST 2002 SRE corpus (left) and its factory noise corrupted (0 dB SNR) counterpart (right). (a) and (b) parametric (all-pole) spectra (c) and (d) non-parametric spectra. The spectra in each plot have been shifted for better visualization.

The **RLP** method [15] introduces a penalty measure into the filter optimization. The objective function for RLP becomes $E = \sum_n (x(n) - \sum_{k=1}^{p} v_k x(n-k))^2 + \lambda g(\mathbf{v})$, where λ is the regularization constant and $g(\mathbf{v})$ is the penalty measure which is a function of prediction coefficients, $\mathbf{v} = [v_0,\ v_1,\ \ldots,\ v_p]^{\mathrm{T}}$. As λ increases, the corresponding spectral envelope gets smoother and as $\lambda \to 0$, RLP reduces to the conventional LP method. Minimizing the regularized objective function leads to the following solution:

$$\mathbf{v}_{\mathrm{opt}}^{\mathrm{RLP}} = -(\mathbf{R} + \lambda \mathbf{DRD})^{-1}\mathbf{r}, \quad (6)$$

where $\mathbf{R}$ is the autocorrelation matrix, $\mathbf{D}$ is a diagonal matrix in which each diagonal element has the value of the row number and $\mathbf{r}$ is the autocorrelation vector. For WLP, SWLP, XLP, SXLP and RLP the spectral envelope is computed by Fourier-transforming the corresponding all-pole transfer function.

The **MVDR** spectrum estimation method (also known as the Capon method or maximum likelihood (ML) spectrum estimation method) [13] was shown to be an effective method that models the unvoiced or mixed speech spectra by using the LP coefficients. An m^{th} order MVDR spectrum is computed by

$$S_{\mathrm{MVDR}}(f) = \frac{1}{|\sum_{k=-m}^{m} \mu(k) e^{-j2\pi fk}|^2}, \quad (7)$$

where m is the MVDR filter order and the parameters $\mu(k)$ are computed by a simple non-iterative method from the LP coefficients [13].

Fig. 1 shows the short-term spectra of an original voiced speech frame (left panel) and its 0 dB noisy counterpart (right panel) computed using different methods. Prediction order $p = 20$ has been used for the all-pole methods.

3. EXPERIMENTAL SETUP

3.1. Corpora, classifier and error measurement

To compare different spectrum estimators, we use a Gaussian mixture model - universal background model (GMM-UBM) [3] with test normalization (Tnorm) applied on the log-likelihood ratio scores.

Table 1. EERs (in %) for different spectrum estimators under additive factory noise (The smallest EER for each SNR level within each sub-group is underlined and globally smallest EER in each row is bolded).

	Baseline methods		Temporally weighted methods				Multitaper methods			Other methods		
SNR (dB)	DFT	LP	WLP	SWLP	XLP	SXLP	Thomson	Multipeak	SWCE	ICS	MVDR	RLP
original	7.65	7.44	7.48	7.81	7.94	7.78	7.39	7.41	**7.32**	8.01	7.62	7.57
20	8.08	7.83	**7.81**	8.22	8.04	7.98	7.95	8.18	8.00	8.45	8.30	7.81
10	9.32	**8.50**	8.79	9.11	8.85	8.85	9.12	9.42	9.20	9.55	9.12	8.75
0	10.46	**9.93**	10.34	10.06	10.01	9.99	10.63	11.07	11.09	10.88	10.36	10.29
-10	15.35	14.96	15.19	**14.35**	14.55	14.73	15.43	15.59	15.26	16.05	14.78	15.02

Table 2. EERs (in %) for different spectrum estimators under additive babble noise (The smallest EER for each SNR level within each sub-group is underlined and globally smallest EER in each row is bolded).

	Baseline methods		Temporally weighted methods				Multitaper methods			Other methods		
SNR (dB)	DFT	LP	WLP	SWLP	XLP	SXLP	Thomson	Multipeak	SWCE	ICS	MVDR	RLP
original	7.65	7.44	7.48	7.81	7.94	7.78	7.39	7.41	**7.32**	8.01	7.62	7.57
20	7.83	7.78	**7.71**	8.11	7.94	7.93	7.76	7.96	7.85	8.28	8.19	7.81
10	8.85	**8.58**	8.70	8.78	8.68	8.85	8.85	9.25	9.00	9.56	9.19	8.92
0	11.62	11.23	11.47	10.93	**10.63**	10.83	11.65	12.19	12.34	11.91	11.70	10.94
-10	21.27	20.35	21.02	19.69	20.35	20.23	21.77	21.86	21.52	22.03	**19.68**	20.12

This choice is mainly motivated by the large number of methods and control parameters to be evaluated. Experiments are conducted on the NIST 2002 SRE corpus which consists of 330 target speakers (139 males, 191 females) and a total number of 39256 trials (2982 genuine, 36277 impostor). The training material consists of 2 minutes of conversational telephone speech while the duration of test utterances varies from 15 to 45 seconds. Gender dependent background and Tnorm models with 512 Gaussians are trained using the NIST 2001 SRE corpus.

For the experiments under additive noise, we use factory and babble noises from the NOISEX-92 database [1]. The target models, background models and Tnorm cohort models are trained using original data and the noise is added to the test samples with a given average signal-to-noise-ratio (SNR). Five different values of SNR are considered in the experiments, $\mathrm{SNR} \in \{\mathrm{clean}, 20, 10, 0, -10\}$ dB where *clean* refers to the original NIST samples. We apply spectral subtraction on the test samples as a preprocessing method.

We use equal error rate (EER) as the performance criterion. EER corresponds to the threshold at which the false alarm rate (P_{fa}) and miss rate (P_{miss}) are equal. Additionally, a few selected detection error trade-off (DET) curves are plotted to analyze the complete behaviour of the methods.

3.2. Feature extraction

MFCC features are extracted from 30 ms Hamming windowed frames every 15 ms. Different methods are evaluated on the magnitude spectrum estimation step of the MFCC extraction procedure. 12 MFCCs are extracted from a 27-channel mel-filterbank. After RASTA filtering, the first and second order derivatives (Δ and $\Delta\Delta$) are appended to the MFCC vectors. The last two steps are energy-based voice activity detection (VAD) and cepstral mean and variance normalization (CMVN).

3.3. Parameter setup for the spectrum estimators

Each spectrum estimator have its own control parameters. The number of spectral bins, 512, is common for all methods. For the all-pole methods, the prediction order is set to $p = 20$. In the temporally weighted LP methods, short-term energy window durations are fixed to be $M = 20$ as in [9, 18]. The MVDR filter order is set to $m = 28$ and for RLP, $\lambda = 10^{-4}$ regularization parameter is used. For the ICS spectrum estimator, $I = 6$ iteration is selected with 30 cepstral smoothing coefficients[1]. $K = 6$ tapers are used for SWCE and multipeak windowing methods and $K = 4$ tapers for Thomson in the multi-taper method.

[1] http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html

4. SPEAKER VERIFICATION RESULTS

Table 1 and Table 2 summarize the results for different spectrum estimators under additive factory and babble noises, respectively. The smallest EER of each row is bolded. Fig. 3 and Fig. 4 display the DET curves for -10 dB SNR level of a few selected methods for factory and babble noise, respectively.

For **original** data we observe that;

- The LP-based methods achieve slightly better recognition rate than the DFT technique (7.65 %). XLP (7.34 %) and SWLP (7.34 %) outperform the other all-pole methods.
- The Multitaper method with the SWCE window outperforms the multipeak and Thomson windows. The SWCE method is the best choice for original data condition.
- ICS technique gives the highest EER of 8.01%.

For **additive factory noise** contamination:

- Conventional LP method gives the smallest EER at high SNR levels (20 dB condition is slightly worse than WLP and RLP).
- Thomson method outperforms SWCE and multipeak techniques for 20, 10 and 0 dB conditions (e.g., 10.63 %, 11.09 % and 11.07 % at 0 dB for Thomson, SWCE and multipeak methods, respectively) but SWCE wins at -10 dB SNR.
- For the noisiest case (-10 dB SNR) SWLP (14.35 %) outperforms the other weighted all-pole methods (15.19 %, 14.55 % and 14.73 % for WLP, XLP and SXLP, respectively).

[1] Note that this is different from the number of MFCCs which is 12 in all methods

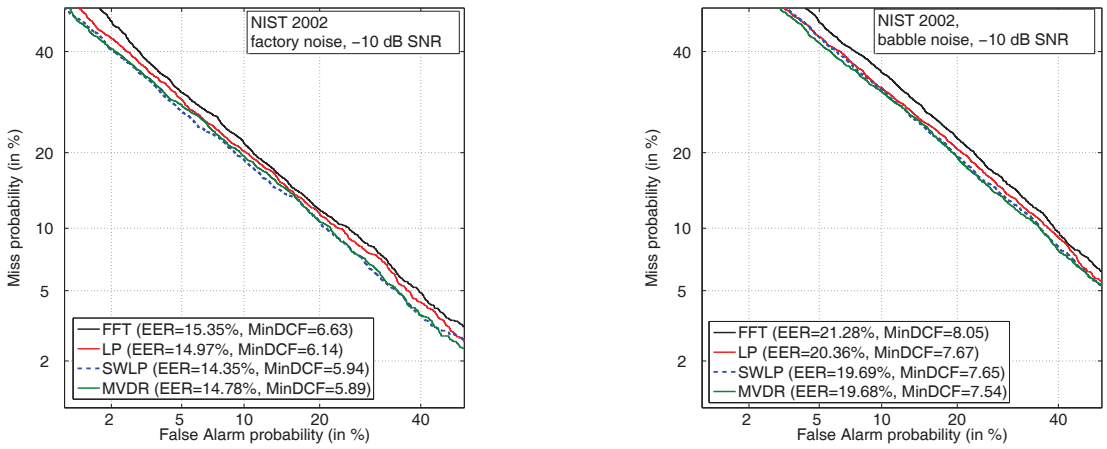

Fig. 2. DET curves for different spectrum estimation methods under additive factory (left) and babble (right) noise (−10 dB SNR level)

- The ICS method gives the highest EERs at nearly all SNR levels.

For **additive babble noise** contamination, the performance of spectrum estimators are very similar to the case of factory noise. However, MVDR method yields the smallest EER for the noisiest case (-10 dB SNR).

5. CONCLUSION

We compared 12 different spectrum estimators in speaker verification under additive noise contamination. In clean condition, multitaper method with SWCE window outperformed remaining methods. In the baseline group, LP outperformed FFT in all cases. WLP yielded smaller EER for high SNR levels in comparison to other temporally weighted methods. However, SWLP gave the smallest EER in the noisiest case. For the multitaper methods, smallest EERs have been obtained with Thomson window for high SNRs. In the noisiest case, the best recognition accuracy has been obtained with SWCE. Under factory noise, in noisiest case the SWLP method showed improvement on recognition accuracy over standard DFT and LP techniques. In our experiments, for babble noise at -10 dB SNR level, SWLP and MVDR techniques are found to be the two best choices. Overall, the spectrum estimation step has a significant impact on recognition performance under additive noise contamination and temporally weighted methods and MVDR technique are promising for speaker recognition.

6. REFERENCES

[1] T. F. Quatieri, *Discrete Time Speech Signal Processing*, Prentice Hall, 2002.

[2] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Comm.*, vol. 52, no. 1, pp. 12–40, Jan. 2010.

[3] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Dig. Sig. Proc.*, vol. 10, no. 1, pp. 19–41, Jan. 2000.

[4] W.M. Campbell, D.E. Sturim, and D.A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Proc. Lett.*, vol. 13, no. 5, pp. 308–311, May 2006.

[5] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumochel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 15, no. 4, pp. 1435–1447, May 2007.

[6] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, 2007.

[7] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech and Audio Proc.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[8] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Dig. Sig. Proc.*, vol. 10, no. 1-3, pp. 42–54, Jan. 2000.

[9] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku, "Temporally weighted linear prediction features for tackling additive noise in speaker verification," *IEEE Signal Proc. Lett.*, vol. 17, no. 6, pp. 599–602, June 2010.

[10] J. Sandberg, M. Hansson-Sandsten, T. Kinnunen, R. Saeidi, P. Flandrin, and P. Brognat, "Multitaper estimation of frequency-warped cepstra with application to speaker verification," *IEEE Signal Proc. Lett.*, vol. 17, no. 4, pp. 343–346, Apr. 2010.

[11] T. Kinnunen, R. Saeidi, J. Sandberg, and M. Hansson-Sandsten, "What else is new than Hamming window? robust MFCCs for speaker recognition via multitapering," in *Interspeech 2010*, Makuhari, Japan, Sept., pp. 2734–3737.

[12] S. Dharanipragada, U. H. Yapanel, and B. D. Rao, "Robust feature extraction for continuous speech recognition using MVDR spectrum estimation method," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 15, no. 1, pp. 224–234, Jan. 2007.

[13] M. N. Murthi and B. D. Rao, "All-pole modeling of speech based on the minimum variance distortionless response spectrum," *IEEE Trans. Speech and Audio Proc.*, vol. 8, no. 3, pp. 221–239, May 2000.

[14] C. Liang, X. Zhang, L. Yang, J. Zhang, and Y. Yan, "Perceptual MVDR-based cepstral coefficients PMCCs for speaker recognition," in *Signal Processing*, Beijing, China, 2010, pp. 2386–2389.

[15] L. A. Ekman, W. B. Kleijn, and M. N. Murthi, "Regularized linear prediction of speech," *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 16, no. 1, pp. 65–73, Jan. 2008.

[16] A. Robel, F. Villavicencio, and X. Rodet, "On cepstral and all-pole based spectral envelope modeling with unknown model order," *Pattern Rec. Lett.*, vol. 28, pp. 1343–1350, 2007.

[17] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, "Stabilised weighted linear prediction," *Speech Comm.*, vol. 51, no. 5, pp. 401–411, 2009.

[18] J. Pohjalainen, R. Saeidi, T. Kinnunen, and P. Alku, "Extended weighted linear prediction (XLP) analysis of speech and its application to speaker verification in adverse conditions," *Interspeech*, vol. II, pp. 1477–1480, 2010.