# Comparison of Two Kinds of Speaker Location Representation for SVM-based Speaker Verification

*Xianyu Zhao* [1], *Yuan Dong* [1, 2], *Hao Yang* [2], *Jian Zhao* [2], *Liang Lu* [2], *Haila Wang* [1]

[1] France Telecom R&D Center (Beijing), Beijing, 100080, P. R. China
{xianyu.zhao, yuan.dong, haila.wang}@orange-ftgroup.com
[2] Beijing University of Posts and Telecommunications, Beijing, 100876, P. R. China
yuandong@bupt.edu.cn {haoyang.hy, michaeljianzhao, luliang07}@gmail.com

## Abstract

In anchor modeling, each speaker utterance is represented as a fixed-length location vector in the space of reference speakers by scoring against a set of anchor models. SVM-based speaker verification systems using the anchor location representation have been studied in previously reported work with promising results. In this paper, linear combination weights in reference speaker weighting (RSW) adaptation are explored as an alternative kind of speaker location representation. And this kind of RSW location representation is compared with the anchor location representation in various speaker verification tasks on the 2006 NIST Speaker Recognition Evaluation corpus. Experimental results indicate that with long utterances for reliable maximum likelihood estimation in RSW, the RSW location representation leads to better speaker verification performance than the anchor location; while the latter is more effective for verification of short utterances in high-dimensional representation space.

**Index Terms**: speaker verification, speaker location, anchor modeling, reference speaker weighting, support vector machines

## 1. Introduction

The idea of using a set of reference speakers for speech modeling has been extensively studied for many tasks in speech processing, e.g., rapid speaker adaptation [1], [2], speaker recognition [3]–[5] and tracking [6]. In [3], anchor models were introduced for speaker verification and indexing, in which speaker utterance data is scored against a set of reference speaker models to determine its corresponding location vector in the space of reference speakers. In [4], [5], it is shown that speaker verification systems using such kind of anchor location representation can achieve state-of-the-art verification performance.

Reference speaker weighting (RSW) was developed for rapid speaker adaptation in speech recognition [1]. It builds models for new speakers as a linear combination of reference speaker models in a maximum likelihood sense. It has been shown that the RSW adapted model may improve speech recognition performance of the speaker-independent model with a small amount of adaptation data. In [7], [8], RSW was also employed for enrolling speakers with limited enrollment data in speaker verification tasks. In this paper, instead of using RSW adapted target speaker models to calculate likelihood scores of each frame in utterance data , the set of linear combination coefficients used in RSW modeling is stacked into a vector, which is then used as a kind of speaker location representation in the space of reference speakers.

Since it appeared in the early nineties as optimal margin classifiers in the context of Vapnik's statistical learning theory [9], support vector machines (SVMs) have recently become one of the most important and widely used classification techniques in the field of speaker recognition [10]–[13]. In this paper, support vector machines (SVMs) are applied to discriminate these speaker location vectors in reference speaker space for speaker verification purpose.

These two kinds of speaker location representation, i.e., the anchor and RSW location respectively, are compared in various verification tasks on the 2006 NIST SRE corpus [14]. In one task, the duration of enrollment and test utterances is limited to about 10 seconds; while in the other task, longer enrollment and test utterances are used (about 2 minutes). Experimental results indicate that for short utterances the anchor location representation obtains better verification performance than the RSW location representation; on the other hand, the latter is more effective for longer utterances.

The rest of the paper is organized as follows. In Section 2, we discuss the derivation of location vector representation for speaker utterance data based on anchor modeling and reference speaker weighting respectively. In Section 3, speaker verification with SVMs is presented. Section 4 evaluates these two kinds of speaker location representation through a series of SVM-based speaker verification experiments on the 2006 NIST Speaker Recognition Evaluation (SRE) corpus. And, Section 5 concludes the paper with a summary and future work.

## 2. Speaker representation by location in the space of reference speakers

In this section, we describe how an utterance and its underlying speaker can be represented as a location vector by its relationship to a predetermined set of reference speakers. Anchor location is derived by scoring the utterance data against a set of reference speaker models, whereas RSW location vector is composed by stacking the linear combination coefficients which are used to construct a model for the utterance from reference speaker models in the RSW adaptation process.

### 2.1. Location representation with anchor modeling

In this approach, speaker location in the space of reference speakers is represented by the following vector, $\mathbf{v}$ , [3]–[6]

$$\mathbf{v} = \left[ \tilde{p}\left(\mathbf{x}\middle|\bar{\lambda}_1\right), \tilde{p}\left(\mathbf{x}\middle|\bar{\lambda}_2\right), \cdots, \tilde{p}\left(\mathbf{x}\middle|\bar{\lambda}_E\right) \right]^T, \quad (1)$$

where $(\cdot)^T$ stands for vector transpose, $\left\{\bar{\lambda}_i; i = 1,2,\cdots,E\right\}$ is a set of well trained reference speaker models (called anchor models), which are modeled as Gaussian Mixture Models

(GMMs) and MAP adapted from a Universal Background Model (UBM) [15] in this study; $\tilde{p}\left(\mathbf{x}\middle|\bar{\lambda}_i\right)$ is the normalized log-likelihood of the speaker utterance data $\mathbf{x}$ (of $T$ acoustic feature vectors) for the $i$-th anchor model, $\bar{\lambda}_i$, relative to the Universal Background Model, $\bar{\lambda}_{UBM}$,

$$\tilde{p}\left(\mathbf{x}\middle|\bar{\lambda}_i\right)=\frac{1}{T}\log\frac{p\left(\mathbf{x}\middle|\bar{\lambda}_i\right)}{p\left(\mathbf{x}\middle|\bar{\lambda}_{UBM}\right)}. \tag{2}$$

### 2.2. Location representation with reference speaker weighting (RSW)

In this case, the reference space, called RSW space in the following discussion, is spanned by supervectors of reference speaker models. In this study, these reference speaker models are MAP adapted GMMs from the UBM as in the anchor case mentioned above; so that all reference speaker models have the same structure (e.g. the number of mixture components) as the UBM. For each reference speaker, a GMM supervector, $\bar{\Lambda}_i$, is formed by concatenating all of the mean vector parameters in corresponding GMM model, $\bar{\lambda}_i$, i.e.

$$\bar{\Lambda}_i=\left[\left(\mu_1^i\right)^T,\left(\mu_2^i\right)^T,\cdots,\left(\mu_K^i\right)^T\right]^T, \tag{3}$$

where $\mu_k^i$ is the mean vector of $k$-th Gaussian component in the $i$-th reference model, $\bar{\lambda}_i$.

For a new utterance, $\mathbf{x}$, a model, $\bar{\lambda}_{\mathbf{x}}$, is constructed for it through reference speaker weighting (RSW) [1]. In this way, the GMM supervector of $\bar{\lambda}_{\mathbf{x}}$, denoted by $\bar{\Lambda}_{\mathbf{x}}$, is represented as a linear combination of the GMM supervectors of reference speaker models, $\left\{\bar{\Lambda}_i;i=1,2,\cdots,E\right\}$, as

$$\bar{\Lambda}_{\mathbf{x}}=\sum_{i=1}^{E}\mathbf{v}(i)\cdot\bar{\Lambda}_i. \tag{4}$$

Then, speaker location in the RSW space is represented by this set of linear combination coefficients, $\left\{\mathbf{v}(i);i=1,2,\cdots,E\right\}$:

$$\mathbf{v}=\left[\mathbf{v}(1),\ \mathbf{v}(2),\ \cdots,\ \mathbf{v}(E)\right]^T. \tag{5}$$

In this paper, the location vector in the RSW space is found in a maximum likelihood sense through

$$\hat{\mathbf{v}}=\arg\max_{\mathbf{v}}P\left(\mathbf{x}\middle|\mathbf{v},\left\{\bar{\Lambda}_i;i=1,2,\cdots,E\right\}\right), \tag{6}$$

where $\mathbf{x}\triangleq\left\{\mathbf{x}_1,\cdots,\mathbf{x}_T\right\}$ represents the utterance data (of $T$ acoustic feature vectors); $P\left(\mathbf{x}\middle|\mathbf{v},\left\{\bar{\Lambda}_i;i=1,2,\cdots,E\right\}\right)$ is the likelihood of utterance data given the location vector and the set of reference speaker models, which is calculated as

$$\begin{aligned}&P\left(\mathbf{x}\middle|\mathbf{v},\left\{\bar{\Lambda}_i;i=1,2,\cdots,E\right\}\right)\\&=\prod_{t=1}^{T}\left\{\sum_{k=1}^{K}w_kP\left(\mathbf{x}_t\middle|k,\mathbf{v},\left\{\bar{\Lambda}_i;i=1,2,\cdots,E\right\}\right)\right\}\\&=\prod_{t=1}^{T}\left\{\sum_{k=1}^{K}w_kN\left(\mathbf{x}_t;\mu_k^x,C_k\right)\right\},\end{aligned} \tag{7}$$

where $w_k$, $\mu_k^x$ and $C_k$ are respectively the weight, mean vector, and covariance matrix of the $k$-th component in the utterance GMM model, $\bar{\lambda}_x$. In this study, the weights and covariance matrices of mixture components are shared among all of the reference speaker models, utterance models and UBM. For the mean vector, $\mu_k^x$, according to (4), we get:

$$\mu_k^x=\sum_{i=1}^{E}\mathbf{v}(i)\mu_k^i. \tag{8}$$

The maximization in (6) is done through the Expectation-Maximization (EM) algorithm [16], i.e., iteratively optimizing an auxiliary function $Q(\mathbf{v},\mathbf{v}')$ with respect to $\mathbf{v}$,

$$\begin{aligned}Q(\mathbf{v},\mathbf{v}')=&\sum_{t=1}^{T}\sum_{k=1}^{K}P\left(k\middle|\mathbf{x}_t,\mathbf{v}',\left\{\bar{\Lambda}_i;i=1,2,\cdots,E\right\}\right)\\&\log P\left(\mathbf{x}_t\middle|k,\mathbf{v},\left\{\bar{\Lambda}_i;i=1,2,\cdots,E\right\}\right),\end{aligned} \tag{9}$$

where $\mathbf{v}'$ is the current estimate of location vector and $P\left(k\middle|\mathbf{x}_t,\mathbf{v}',\left\{\bar{\Lambda}_i;i=1,2,\cdots,E\right\}\right)$ is the *posteriori* probability of the $k$-th Gaussian component given the utterance data $\mathbf{x}_t$ and the current location estimate $\mathbf{v}'$,

$$\begin{aligned}&P\left(k\middle|\mathbf{x}_t,\mathbf{v}',\left\{\bar{\Lambda}_i;i=1,2,\cdots,E\right\}\right)=\\&\frac{w_kP\left(\mathbf{x}_t\middle|k,\mathbf{v}',\left\{\bar{\Lambda}_i;i=1,2,\cdots,E\right\}\right)}{\sum_{k'=1}^{K}w_{k'}P\left(\mathbf{x}_t\middle|k',\mathbf{v}',\left\{\bar{\Lambda}_i;i=1,2,\cdots,E\right\}\right)}.\end{aligned} \tag{10}$$

Let $\partial Q/\partial\mathbf{v}(i)=0,i=1,2,\cdots,E$, we obtain the update equation for each $\mathbf{v}(i)$, $i=1,2,\cdots,E$:

$$\begin{aligned}&\sum_{t=1}^{T}\sum_{k=1}^{K}P\left(k\middle|\mathbf{x}_t,\mathbf{v}',\left\{\bar{\Lambda}_i;i=1,2,\cdots,E\right\}\right)\left(\mu_k^i\right)^T\Sigma_k^{-1}\mathbf{x}_t\\&=\sum_{l=1}^{E}\mathbf{v}(l)\cdot\left[\sum_{t=1}^{T}\sum_{k=1}^{K}P\left(k\middle|\mathbf{x}_t,\mathbf{v}',\left\{\bar{\Lambda}_i;i=1,2,\cdots,E\right\}\right)\left(\mu_k^i\right)^T\Sigma_k^{-1}\mu_k^l\right]\end{aligned} \tag{11}$$

## 3. SVM-based speaker verification by location in the space of reference speakers

For SVM-based speaker verification by location in the space of reference speakers, the location vector $\mathbf{v}$ is treated as input feature and modeled using support vector machines. In the standard formulation, an SVM, $f(\mathbf{v})$, is given by

$$f(\mathbf{v})=\sum_{i=1}^{M}\alpha_ik\left(\bar{\mathbf{v}}_i,\mathbf{v}\right)+b, \tag{12}$$

where $k\left(\mathbf{v}_1,\mathbf{v}_2\right)$ is a kernel function. In this study, linear kernel is used, i.e.

$$k\left(\mathbf{v}_1,\mathbf{v}_2\right)=\left(\mathbf{v}_1\right)^T\mathbf{v}_2. \tag{13}$$

The parameters, $b$ and $\left\{\alpha_i,\bar{\mathbf{v}}_i;i=1,\cdots,M\right\}$, are obtained through a training process that maximizes the margin between two classes (positive vs. negative). In this study, SVMTorch is used as SVM trainer [17]. For classification, a decision is made upon whether the value, $f(\mathbf{v})$, is above or below a threshold.

In the application of SVMs for speaker verification, an SVM is trained for each target speaker using the location vectors of the speaker's enrollment utterances as positive examples, and the location vectors of all utterances from background speakers in some development data as negative examples.

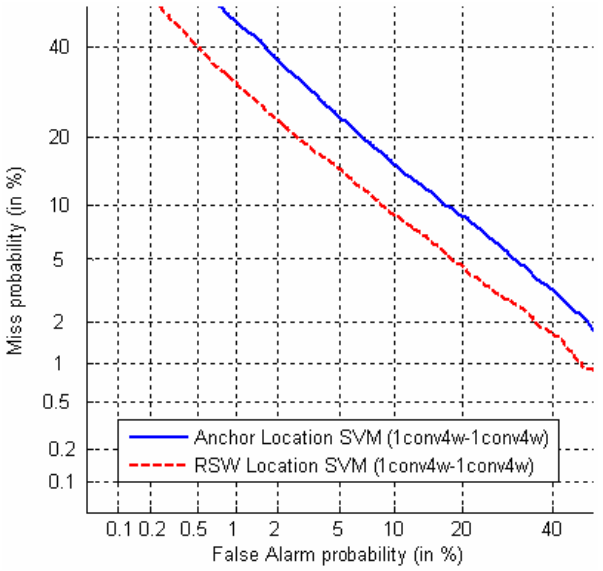


Figure 1: *DET curves for SVM-based speaker verification using the anchor and RSW location respectively in the 2006 NIST SRE 1conv4w-1conv4w task. 583 reference speaker models are used in the representation space.*

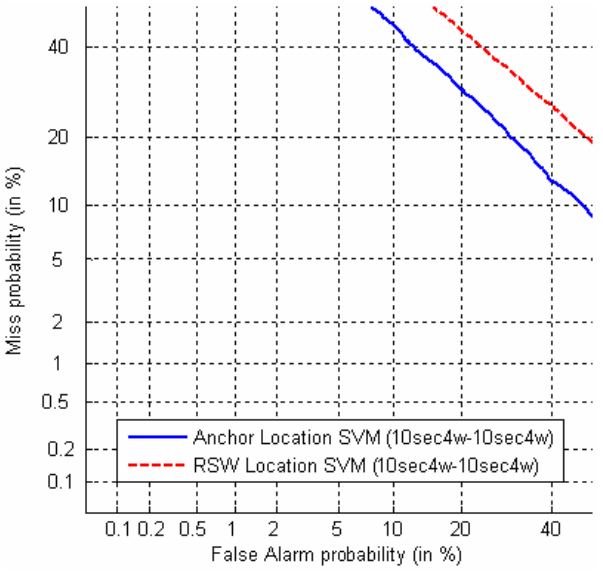


Figure 2: *DET curves for SVM-based speaker verification using the anchor and RSW location respectively in the 2006 NIST SRE 10sec4w-10sec4w task. 583 reference speaker models are used in the representation space.*

## 4. Experimental results

In this section, we report speaker verification experiments by location in the space of reference speakers. Section 4.1 presents some general experimental setup information about task, corpora and acoustic features used. The results of these experiments are discussed in Section 4.2.

### 4.1. Experimental setup

For cepstral feature extraction, a 13-dimensional PLP is calculated every 10 ms using a 25ms Hamming window. First, second and third order derivatives over a $\pm 2$ frame span are computed and appended to each feature vector, which results in dimensionality 52. Heteroscedastic linear discriminant analysis (HLDA) is then used to decorrelate the features and to reduce the dimensionality from 52 to 51 (1 dimension is left out as nuisance). RASTA, feature mapping and histogram equalization (HEQ) are applied to improve channel and noise robustness of cepstral features. A gender independent UBM with 2048 Gaussians is trained using about 40 hours of data from the Switchboard corpora (I, II and Cellular parts). And, in the speech data for UBM training, there are 583 speakers (248 male and 335 female speakers). Their models are MAP adapted from the UBM and used as reference speaker models in anchor modeling and RSW. The relevance factor in MAP adaptation is set to be 16 (only the means are adapted).

Speaker verification experiments were conducted on the 2006 NIST SRE corpus [14]. The anchor and RSW location representation are compared in two tasks. The first task uses one conversation side for training and testing (denoted as 1conv4w-1conv4w [14], the duration of pure speech after voice activity detection is about 2 minutes). This task involves 608 speakers, 3,612 true trials and 47,836 false trials. The other task uses conversation excerpts of about 10 seconds for training and testing (denoted as 10sec4w-10sec4w [14]), which involves 2,942 true trials and 29,608 false trials. Location vectors of utterances in the 2004 NIST SRE dataset are calculated as negative examples in SVM training.

Results are presented using Detection Error Tradeoff (DET) plots. Along with Equal Error Rate (EER), the minimum detection cost function (DCF) value, as defined by NIST [14], is also used as an overall performance measure.

### 4.2. Results

Figure 1 shows the DET curves for SVM-based speaker verification using the anchor and RSW location representation respectively in the 2006 NIST SRE 1conv4w-1conv4w task. All 583 reference models are used to construct the representation space for location vectors. Similarly, Figure 2 compares the DET curves for these two kinds of location representation in the 2006 NIST SRE 10sec4w-10sec4w task.

We can see that with long enrollment and test utterances in the 1conv4w-1conv4w task, the RSW location representation leads to better verification performance than the anchor location representation. While for short utterances in the 10sec4w-10sec4w task, verification using the anchor location representation is more effective.

This may be related with the fact that the RSW location vectors are derived as parameters of a model adaptation process in a maximum likelihood sense. With enough data to guarantee reliable maximum likelihood estimation in equation (11), the model-based representation of utterance data in RSW is capable of reducing some noisy effects in original utterances and derives more effective location representation in the space of reference speakers. However, for short utterances, it may not be able to carry out the maximum likelihood estimation in RSW reliably with respect to a large number of reference speaker models, which would affect the precision of derived RSW location representation and degrade following speaker verification performance. For the above mentioned experiments in the 10sec4w-10sec4w task, there are about 1,000 frames of acoustic features in an utterance of about 10 seconds; yet, there are 583 parameters needed to be

Table 1. *SVM-based speaker verification results for the anchor and RSW location representation in the 2006 NIST SRE 10sec4w-10sec4w task. They are compared when varying the number of reference speaker models in the representation space.*

| Dim. of Ref. Space | Anchor Location | | RSW Location | |
|---|---|---|---|---|
| | EER (%) | DCF (x100) | EER (%) | DCF (x100) |
| 100 | 28.11 | 9.47 | 28.76 | 9.48 |
| 200 | 26.68 | 9.17 | 27.57 | 9.24 |
| 300 | 25.46 | 9.05 | 27.97 | 9.29 |
| 400 | 24.92 | 8.97 | 28.45 | 9.31 |
| 500 | 24.75 | 8.92 | 30.59 | 9.60 |
| 583 | 24.75 | 8.88 | 32.09 | 9.75 |

estimated in RSW to derive a location vector in the representation space having 583 reference speaker models.

Hence, in the following experiments, we vary the number of reference speaker models in the representation space to investigate whether this would help the parameter estimation process in RSW and to study how this would affect verification performance. Table 1 summarizes the EER and the minimum NIST DCF values. Subsets of reference speakers are selected from all of the 583 speakers using the method proposed in [4], which selects a subset of reference voices that covers the main variability of initial set of reference speakers.

Table 1 shows that for the anchor location representation, verification performance improves as more reference speaker models are used. This is due to the fact that each component in the anchor location vectors is estimated independently with other components by scoring utterance data against corresponding reference speaker model. Adding more reference speaker models would not affect the estimation precision of each component; and more reference speaker models would to some extent inject more information into the location vectors, which could help SVMs to discriminate speakers in the space of reference speakers.

For the RSW location representation, from Table 1 we can see that best verification performance is achieved with 200 reference speaker models. Although more reference speaker models would result in more informative representation space, it would also cause parameter estimation problem in the maximum likelihood formulation of RSW. Hence compared with using all 583 reference speaker models, using less reference speakers could guarantee more robust parameter estimation in RSW, which in turn improves verification performance.

## 5. Conclusions

In this study, two different types of speaker location representation based on anchor modeling and reference speaker weighting (RSW) respectively are compared for SVM-based text-independent speaker verification. Experimental results show that with long utterances to guarantee enough data for reliable maximum likelihood parameter estimation in RSW, the RSW location vector could lead to more effective speaker location representation and obtain better verification performance than the anchor location representation. However, for short utterances (e.g. about 10 seconds), verification system using the anchor location representation outperforms that using the RSW location. This is related with the fact that the deficiency of data would cause unreliable maximum likelihood estimation of the RSW location vectors in a high-dimensional reference speaker space and degrade following speaker verification performance. This justifies a need to balance the complexity of location representation against the amount of available utterance data, which will be studied further in future work.

## 6. References


[1] T. Hazen, "The use of speaker correlation information for automatic speech recognition," Ph.D. Thesis, MIT, Cambridge, Jan. 1998.

[2] R. Kuhn, J.-C. Junqua, P. Nguyen and N. Niedzielski, "Rapid speaker adaptation in Eigenvoice space," *IEEE Trans. Speech and Audio Processing*, vol.8, no.6, pp. 695-707, Nov. 2000.

[3] D. Sturim, D. Reynolds, E. Singer, and J. P. Campbell, "Speaker indexing in large audio database using anchor models," in *Proc. ICASSP*, 2001, pp. 429-432, 2001.

[4] Y. Mami, D. Charlet, "Speaker recognition by location in the space of reference speakers," *Speech Communication*, vol.48, pp. 127-141, 2006.

[5] X. Zhao, Y. Dong, H. Yang, J. Zhao and H. Wang, "SVM-based speaker verification by location in the space of reference speakers," in *Proc. ICASSP*, 2007.

[6] M. Collet, D. Charlet and F. Bimbot, "Speaker tracking by anchor models using speaker segment cluster information," in *Proc. ICASSP*, 2006, pp. 1009-1012.

[7] Man-Wai Mak, R. Hsiao, B. Mak, "A comparison of various adaptation methods for speaker verification with limited enrollment data," in *Proc. ICASSP*, 2006, pp. 929-932.

[8] X. Zhao, Y. Dong, J. Luo, H. Yang and H. Wang, "Multigrained model adaptation with MAP and reference speaker weighting for text independent speaker verification," in *Proc. ICASSP*, 2006, pp. 913-916.

[9] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[10] W. M. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition," in *Proc. ICASSP*, 2002, pp. 161-164.

[11] W. M. Campbell, D. E. Sturim, D. A. Reynolds and A. Solomonoff, "SVM-based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. ICASSP*, 2006, pp. 97-100.

[12] A. O. Hatch, A. Stolcke, "Generalized linear kernels for one-versus-all classification: application to speaker recognition," in *Proc. ICASSP*, 2006, pp. 585-588.

[13] V. Wan, S. Renals, "Speaker verification using sequence discriminant support vector machines," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 2, pp. 203-210, 2005.

[14] "The NIST 2006 speaker recognition evaluation plan," http://www.nist.gov/speech/tests/spk/spk/2006.

[15] D. Reynolds, T. Quatieri and R. Dunn, "Speaker verification using adapted Gaussian Mixture Models," *Digital Signal Processing*, vol.10, pp. 19-41, 2000.

[16] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1-38, 1977.

[17] R. Collobert, S. Bengio, "SVMTorch: Support vector machines for large–scale regression problems," *Journal of Machine Learning Research*, vol. 1, pp. 143-160, 2001.