NOISE ROBUST EXEMPLAR-BASED CONNECTED DIGIT RECOGNITION

Jort F. Gemmeke[1], *Tuomas Virtanen*[2]

[1]Department of Linguistics, Radboud University, Nijmegen, The Netherlands.
[2]Department of Signal Processing, Tampere University of Technology, Finland.

j.gemmeke@let.ru.nl tuomas.virtanen@tut.fi

ABSTRACT

This paper proposes a noise robust exemplar-based speech recognition system where noisy speech is modeled as a linear combination of a set of speech and noise exemplars. The method works by finding a small number of labeled exemplars in a very large collection of speech and noise exemplars that jointly approximate the observed speech signal. We represent the exemplars using mel-energies, which allows modeling the summation of speech and noise, and estimate the activations of the exemplars by minimizing the generalized Kullback-Leibler divergence between the observations and the model. The activations of the speech exemplars are directly being used for recognition. This approach proves to be promising, achieving up to 55.8% accuracy at signal-to-noise ratio $-5\,$dB on the AURORA-2 connected digit recognition task.

Index Terms**—** Speech recognition, exemplar-based, noise robustness, non-negative matrix factorization, sparsity

1. INTRODUCTION

For the last 30 years Automatic Speech Recognition (ASR) has been completely dominated by the use of Hidden Markov Models (HMMs) [1]. HMM-based ASR performance, however, degrades substantially when speech is corrupted by background noise not seen during training. Additionally, it has become clear that not all speech phenomena can be covered in the form of HMMs. There is a general agreement in the speech community about the need for novel approaches for handling phenomena that HMMs do not account for (cf. [2] and the references therein).

One of these approaches, exemplar-based speech recognition, is based on a psycholinguistic theory which states that mental representations of speech include a record of detail of actual speech signals (called *episodes* or *exemplars*). These exemplars may be be linked to information about idiosyncrasies of the speaker and possibly even the context in which an utterance was produced [3]. Recent advances in computational power have led to an renewed interest in the exemplar-based ASR [1; 4].

In [5; 6] a new approach to exemplar-based speech recognition was introduced. The approach, dubbed *sparse classification (SC)*, is based on the idea that speech signals can be represented as a linear combination of a small set of suitably selected exemplars. The classification is done by finding the smallest number of labeled exemplars in a very large collection of exemplars that *jointly* approximate the observed speech signal. Because there is no need for these exemplars to be close to each other in the original space, SC differs from other exemplar-based approaches to speech recognition, which invariably search for exemplars with the smallest distance to the observed speech signal.

In [5] it was shown that SC can be made noise robust using a missing data technique (MDT) [7]: prior to decoding it is estimated which spectro-temporal elements of the acoustic representations are dominated by speech ('reliable') and which are dominated by background noise ('unreliable' or 'missing'). Decoding is then done by using only the reliable speech features, disregarding the unreliable ones. While this approach proved to be very effective when the identity of reliable features were accurately estimated, the weakness of the method is that it proved sensitive to mask estimation errors.

In this paper, we propose a novel approach to noise robust exemplar-based speech recognition in which noisy speech is modeled by a linear combination of speech and noise exemplars and the activations of the speech exemplars are directly used for recognition. All speech exemplar are associated with one or more state labels; thus, the linear combination of exemplars that represents the noisy speech results in a vector of state activation scores. The activations of the noise exemplars are ignored. This approach avoids having to estimate which feature are unreliable, and recognition is done by finding the optimal sequence of states using a conventional Viterbi decoding backend.

In order to find the linear combination of exemplars we apply non-negative matrix factorization (NMF) which previously has been used for source separation [8; 9; 10]. The main difference with previous NMF approaches is that our approach allows us to do speech decoding directly on exemplar activations, thus avoiding the impact of the otherwise unavoidable reconstruction errors. Another difference is that in many approaches the feature vectors used to represent the noisy speech observations are determined online [11; 10] while we determine these in advance and keep them fixed during decoding. This is somewhat similar to the supervised NMF approach in which NMF [8] or mixture modeling [9] are used to train the feature vectors in advance.

As in [6], we employ a sliding time window approach which allows the exemplars to span multiple time frames. In previous studies it was found that increasing the window length had a large influence on the recognition accuracy. The reason for this is that including more time context constrains the search for suitable exemplars, thus increasing accuracy. It seems likely that this effect will only be greater at lower signal-to-noise ratios (SNR's) since the inclusion of more time context should reduce possible confusions with the (less structured) noise exemplars. Using the connected digit recognition task AURORA-2 , we explore the influence of window size and SNR.

The rest of the paper is organized as follows. In Section 2 we introduce our noisy speech representation model and describe the sliding window approach. In Section 3 we describe how we retrieve the linear combination of exemplars used to represent speech. In

The research of Jort Gemmeke was carried out in the MIDAS project, granted under the Dutch-Flemish STEVIN program. The research of Tuomas Virtanen has been funded by the Academy of Finland.

Section 4 we explain how the exemplar activations are used to do speech recognition. In Section 5 we investigate recognition accuracy as a function of window size and SNR. We present the results in Section 6 and we give our conclusions and plans for future work in Section 7.

2. MODEL FOR NOISY SPEECH

2.1. A sparse representation of noisy speech

In ASR speech signals are represented as a spectro-temporal distribution of acoustic energy, called a power spectrogram, which is represented as a $K \times T$ dimensional matrix (with K frequency bands and T time frames). The power spectrogram of noisy speech is (approximately) equal to the sum of the underlying clean speech and noise power spectrograms. We use this property to directly model noisy speech as a linear combination of speech and noise exemplars. Unlike in most studies, where a log-power representation of speech is used, we use a linear-power representation to ensure additivity of noise and speech exemplars. We express the spectrogram $\mathbf{Y}$ as a single column vector $\boldsymbol{y}$ of length $D = K \cdot T$ by concatenating T subsequent time frames.

We use a training corpus to create a dictionary $\boldsymbol{s}_m$, $(1 \leq m \leq S)$ of speech exemplars, which are speech spectrograms reshaped into vectors as above. Matrix $\mathbf{S}$ is formed as $\mathbf{S} = [\boldsymbol{s}_1 \;\boldsymbol{s}_2 \ldots \boldsymbol{s}_S]$. Similarly, a set of N noise exemplar vectors is used to form matrix $\mathbf{N}$ where each column corresponds to an individual exemplar. The speech and noise exemplars are concatenated so as to form a single dictionary $\mathbf{A} = [\mathbf{S}\;\mathbf{N}]$. The matrix $\mathbf{A}$ has dimensionality $D \times M$, where $M = S + N$. The columns of $\mathbf{A}$ are denoted as $\boldsymbol{a}_m$, $1 \leq m \leq M$.

We assume that each vector $\boldsymbol{y}$, reshaped from a noisy speech spectrogram $\mathbf{Y}$, can be expressed as a linear, non-negative combination of exemplars $\boldsymbol{a}_m$:

$$\boldsymbol{y} = \sum_{m=1}^{M} x_m \boldsymbol{a_m} = \mathbf{A}\boldsymbol{x} \quad \text{s.t.} \quad \boldsymbol{x} \geq 0 \tag{1}$$

with $\boldsymbol{x}$ a M-dimensional activation vector. The non-zero entries in $\boldsymbol{x}$ that correspond to speech exemplars, and the state labels associated with these exemplars, carry the information required to decode the words in the speech, as explained in Section 4.

The use of noise exemplars enables finding a representation where the acoustic mismatch caused by noise is compensated. We include several types in the dictionary of noise exemplars in order not to limit ourselves to speech contaminated by a specific noise type (described in Section 5). The magnitude of noise exemplar activations allows matching different noise levels.

As in [12], we require that $\boldsymbol{x}$ is *sparse*, i.e. most of the weights should be equal to zero so that the noisy speech becomes represented as a combination of a small set of exemplars. The algorithm for finding the value of $\boldsymbol{x}$ is explained in Section 3.

2.2. Sliding window approach for time-continuity

In order to decode utterances of arbitrary lengths, we adopt a sliding time window approach as in [6]. Using this approach we do not have to pre-segment the utterance, and avoid problems at segment boundaries. Consider a speech utterance $\mathbf{U}$ represented as a spectrogram of size $K \times I$. We slide a window $\mathbf{Y}$ spanning T frames through $\mathbf{U}$, with shifts of Δ frames. For each position of $\mathbf{Y}$ in $\mathbf{U}$, we express the spectrogram $\mathbf{Y}$ as a single vector $\boldsymbol{y}$ as described above. The number of windows needed for processing the entire speech signal $\mathbf{U}$ is given by $W = \text{ceil}((I - T)/\Delta) + 1$.

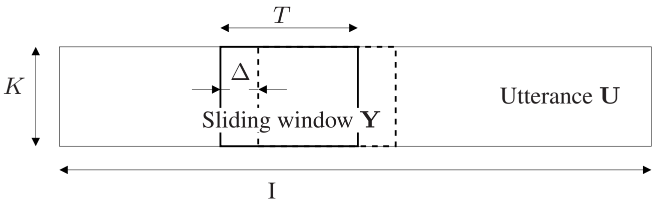

Fig. 1: Schematic diagram of time-continuous classification using overlapping windows.

The utterance $\mathbf{U}$ is now represented by the W subsequent windows: vectors $\boldsymbol{y}_w$, $1 \leq w \leq W$. Concatenating these we form a new matrix $\mathbf{\Psi}$ of dimensions $D \times W$. The ratio of Δ and T determines the degree with which subsequent windows overlap. Larger step sizes Δ reduce computational effort but might decrease recognition accuracy. In this paper we keep the shift constant at $\Delta = 1$ frame.

Using the notation introduced above, we write (1) compactly as:

$$\mathbf{\Psi} = \mathbf{A}\mathbf{X} \quad \text{s.t.} \quad \mathbf{X} \geq 0 \tag{2}$$

with the matrix $\mathbf{X} = [\boldsymbol{x}_1 \;\boldsymbol{x}_2 \ldots \boldsymbol{x}_W]$. $\mathbf{X}$ now describes the activations of the exemplars for the entire utterance.

3. FINDING THE ACTIVATIONS

In order to obtain $\mathbf{X}$ we look for values which are able to represent the noisy speech $\mathbf{\Psi}$ with the model $\mathbf{AX}$, while using a small number of non-negative entries in $\mathbf{X}$. A practical solution is to minimize a cost function

$$d(\mathbf{\Psi}, \mathbf{AX}) + \sum_{m=1}^{M}\sum_{w=1}^{W} [\mathbf{X}.*\mathbf{\Lambda}]_{m,w} \quad \text{s.t.,} \quad \mathbf{X} \geq 0 \tag{3}$$

with $.*$ denoting element-wise multiplication. The first term measures the distance between the noisy observation and the model using function d. The second term penalizes non-zero entries of $\mathbf{X}$ weighted by matrix $\mathbf{\Lambda}$, controlling the degree of sparseness of resulting $\mathbf{X}$. Only non-negative entries of $\mathbf{X}$ are allowed, since a negative activation would correspond to subtracting a power spectrogram, which is not physically realistic. Unlike in most studies, where a single scalar weight is used to penalize all non-zero entries equally, we allow different weights for different types of exemplars in the dictionary. In our experiments enforcing sparseness of noise exemplars was not found to increase the recognition accuracy, while the sparseness of speech exemplars was found to be very important.

As a distance measure we use the generalized Kullback-Leibler (KL) divergence

$$d(\mathbf{\Psi}, \hat{\mathbf{\Psi}}) = \sum_{d=1}^{D}\sum_{w=1}^{W} \mathbf{\Psi}_{d,w} \log(\mathbf{\Psi}_{d,w}/\hat{\mathbf{\Psi}}_{d,w}) - \mathbf{\Psi}_{d,w} + \hat{\mathbf{\Psi}}_{d,w} \tag{4}$$

which has been found to produce good results in sound source separation [11]. Our previous studies [5; 6] used the squared Euclidean distance between the vectors for a log-power feature representation. Since in this study we use linear-power feature representation, the

KL divergence reflects the distribution of natural speech and noise energies better.

The cost function (3) is minimized by first initializing the activations $\mathbf{X}$ to unity, and then iteratively applying the update rule

$$\mathbf{X} \leftarrow \mathbf{X}.*(\mathbf{A}^{\mathrm{T}}(\mathbf{\Psi}./(\mathbf{AX}))))./(\mathbf{A}^{\mathrm{T}}\mathbf{1}+\mathbf{\Lambda}). \tag{5}$$

with $./$ denoting element-wise division and $\mathbf{1}$ an all-one matrix having dimensions $D \times W$.

The cost function (3) is non-increasing under the update rule. This can be proven as in [13] with the Gamma prior having scale 1 and shape $1/[\mathbf{\Lambda}]_{m,w}$.

4. CLASSIFICATION

Each exemplar in the speech part of the dictionary $\mathbf{A}$ is labeled using HMM-state labels obtained from a conventional HMM-based decoder. Exemplars that contain multiple time frames may be associated to more than one state. Using a frame-by-frame state description of the training data used to construct the dictionary, we associate every exemplar $\boldsymbol{s}_m$ with a label vector $\boldsymbol{l}_m$. Denoting the total number of state labels with Q, $\boldsymbol{l}_m$ is a vector of length Q of which the nonzero elements indicate the number of frames in that exemplar that are associated with the states $q \in Q$.

We obtain a label matrix $\mathbf{L}$ of dimensions $Q \times S$ by concatenating all exemplar labels $\boldsymbol{l}_m$: $\mathbf{L} = [\boldsymbol{l}_1 \;\; \boldsymbol{l}_2 \ldots \;\; \boldsymbol{l}_S]$. Using only the part of the activation matrix $\mathbf{X}$ which pertains to speech exemplar activations, denoted $\mathbf{X}_{\mathrm{s}}$, we can now map the observed speech to state likelihoods using:

$$\mathcal{L} = \mathbf{L}\mathbf{X}_{\mathrm{s}} \tag{6}$$

with $\mathcal{L}$ a state-likelihood matrix of dimensions $Q \times W$. The values in $\mathcal{L}$ are normalized between zero and unity.

In the current implementation, silence states are not activated if there is no speech activity, since the NMF algorithm does not select exemplars containing only silence. In order to reduce the number of insertion errors caused by spurious non-silence state activations during silence, it was found beneficial to increase the state likelihoods pertaining to silence.

In our approach, we increase the silence states likelihood as a function of the speech activity. We measure the speech activity in window w by $\alpha_w = \sum_{m=1}^{S} [\mathbf{X}_{\mathrm{s}}]_{m,w}$, and add term $\gamma(1 - \alpha_w / \max_{w'}\{\alpha_{w'}\})$ to the silence state likelihoods in window w. The parameter γ is an empirically determined constant.

Finally, as in [6] we decode the speech utterance by using a Viterbi search for the state sequences which maximize likelihood.

5. EXPERIMENTS

For our recognition experiments we used test set 'A' and 'B' of the AURORA-2 corpus [14]. Test set 'A' comprises 1 clean and 24 noisy subsets, containing four noise types (subway, car, babble, exhibition hall) at six SNR values, $20, 15, 10, 5, 0$ and -5 dB. Test set 'B' contains four different noise types (restaurant, street, airport, train station). Each subset contains 1001 utterances with one to seven digits '0-9' or 'oh'. To reduce computation times, we used a random, representative subset of 10% of the utterances (i.e. 400 utterances per SNR level). Acoustic feature vectors consisted of mel frequency power spectra, spanning $K = 23$ bands, a frame shift of 10ms and a frame length of 25ms.

We created a dictionary of 4000 noise and 4000 clean speech exemplars by randomly selecting windows from the noise and clean speech in the multicondition training set. The multicondition training set of AURORA-2 contains 8440 utterances with the same noises as in test set 'A', at SNR values SNR $= 20, 15, 10, 5$ dB. We repeated the random selection for 4 window lengths, $T \in \{5, 10, 20, 30\}$ frames. The spectrograms were reshaped to vectors and subsequently added as the columns of the dictionary $\mathbf{A}$ as described in Section 2.1. The dictionary $\mathbf{A}$ was normalized by fixing the Euclidean norm to unity along both dimensions. Finally, each observation $\mathbf{\Psi}$ was scaled using the normalization matrices applied to $\mathbf{A}$.

HMM-state based labels of the exemplars were obtained via a forced alignment with the orthographic transcription using the HMM-based recognizer described in [15]. Digits were described by 16 states with an additional 3-state silence word, resulting in a $Q = 179$ dimensional state-space. The rows of the label matrix $\mathbf{L}$ were normalized to have Euclidean unit norm.

The speech decoding system was implemented in MATLAB. The NMF update rule (5) was run for 200 iterations. The optimal value for the sparsity parameter $\mathbf{\Lambda}$ for speech exemplars was determined by maximizing recognition accuracy on a random subset of 250 utterances of the multicondition training database and set to 0.65. $\mathbf{\Lambda}$ was set to zero for noise exemplars. Likewise, the silence state boost parameter γ was determined using the same subset and set to 0.005. Viterbi decoding was done using the backend of the HMM-based decoder described in [15]. That same decoder, which can optionally perform MDT noise-robust decoding, was used for our baseline recognition experiments. When using MDT, the decoder replaces unreliable features with Gaussian-conditioned clean speech estimates using a realistic, binary missing data mask [15].

We carried out recognition experiments with a number of window lengths: $T \in \{5, 10, 20, 30\}$ frames. Recognition accuracies were averaged over the four noise types at each SNR level. To keep correspondence with noise robust ASR research using the AURORA-2 database, we also present the average recognition accuracy over the SNR range 20 to 0 dB.

6. RESULTS AND DISCUSSION

From the recognition accuracies in Tables 2a and 2b it can be observed that the best results for clean speech are achieved using a window length of $T = 10$ frames. This is in correspondence with the results obtained in [6], where we employed a log-power representation. The best accuracy achieved on clean speech is 95.5%, which is slightly lower than the 96.6% accuracy observed in [6] when using 4000 speech exemplars. While performance seems quite disappointing, it should be noted that in [6], we reached up to 98.2% accuracy when using 16000 exemplars. It is likely the recognition accuracy will improve when using more speech exemplars. Further research is necessary to explore the influence of feature representation (linear-power rather than log-power features) and distance measure (Euclidean distance vs KL divergence).

In Table 2a we observe that the best recognition accuracy obtained at SNR -5 dB is 55.8% for test set 'A'. With the baseline method only achieving 17.1% accuracy, this shows that the proposed method is quite noise robust. The reason for this large difference is that our approach does not suffer from errors in the estimation of which features are unreliable and errors in the reconstruction of these features. Moreover, the average accuracy over the SNR range $20 - 0$ dB is competitive with that of the baseline, even though the accuracy at high SNR's is substantially lower.

From Tables 2a and 2b it can be inferred that longer window lengths are found to be optimal at lower SNR's. The reason for this is that including more time context prevents confusion with noise ex-

Table 1: Word recognition accuracy for several window lengths and SNR's.

SNR [dB]	clean	20	15	10	5	0	-5	Avg_{0-20}
baseline	99.7	97.9	95.5	91.4	82.6	62.1	17.1	85.9
T=5	88.7	84.9	81.4	73.9	60.5	38.6	20.9	67.9
T=10	95.5	93.8	92.7	90.2	83.8	69.5	41.0	86.0
T=20	93.5	92.3	91.9	88.8	83.8	72.0	49.3	85.8
T=30	89.5	88.4	88.0	85.5	82.6	74.9	55.8	83.9

(a) Test set 'A'

SNR [dB]	clean	20	15	10	5	0	-5	Avg_{0-20}
baseline	99.7	95.3	91.2	84.3	70.4	40.2	12.2	76.3
T=5	88.7	85.8	83.7	75.6	63.1	40.7	17.3	69.8
T=10	95.5	93.7	90.4	84.6	73.5	50.6	21.2	78.5
T=20	93.5	91.6	88.6	80.8	69.1	45.1	23.3	75.0
T=30	89.5	87.2	85.2	80.4	71.8	54.8	32.4	75.9

(b) Test set 'B'

emplars. At the same time, accuracy at higher SNR's decreases when using longer time windows due to an increased number of deletions: it becomes more difficult to recognize digits with a duration much smaller than the length of the exemplar. This suggests a need for a decoding approach in which multiple window lengths are combined.

When studying the results on test set 'B', displayed in Table 2b, we can observe that recognition accuracy, while still higher than the baseline results for the lower SNR's, drops faster as a function of SNR than in test set 'A'. An obvious explanation would be that this is caused by the mismatch between the noises in the dictionary and the noises observed in test set 'B'. However, noise characteristics may also be play a role, since we observe a similar drop in accuracy for the baseline method which does not make explicit assumptions about the corrupting noise.

It is likely that the noise robustness of our approach can be further improved by reducing the mismatch between the noises in the noise dictionary and those observed in the noisy speech exemplars. This can be done by creating a much larger collection of noise types in the dictionary which will reduce the risk of noise mismatch. Furthermore, the flexibility of the approach allows for extending and updating of the noise dictionary, even during decoding.

7. CONCLUSIONS AND FUTURE WORK

We proposed a novel approach to noise robust exemplar-based speech recognition in which noisy speech is modeled by a linear combination of speech and noise exemplars, with the activations of the speech exemplars directly being used for recognition.

This approach proved to be promising, achieving a substantial improvement over baseline recognition accuracies at lower SNR's.The approach is straightforward, flexible and easy to implement. As such, it can serve as a platform for new research in exemplar-based speech recognition, and noise robust speech recognition in particular.

Future work will focus on improving clean speech accuracy. Several approaches are possible, such as adapting the speech dictionary to the speaker during decoding, or using different cost functions which emphasize the small-scale structure normally emphasized by using the log-power feature representation. Additionally, it is conceivable that better results can be obtained using a more informed, non-random, sampling method to construct the noise and speech dictionary.

8. REFERENCES

[1] H. Bourlard, H. Hermansky, and N. Morgan, "Towards increasing speech recognition error rates," *Speech Communication*, vol. 18, pp. 205–231, 1996.

[2] L. Deng and H. Strik, "Structure-based and template-based automatic speech recognition - comparing parametric and non-parametric approaches," in *Proc. of INTERSPEECH*, 2007, pp. 898–901.

[3] J. Goldinger, "Echoes of echoes? an episodic theory of lexical access," *Psychological Review*, vol. 105, pp. 251–279, 1998.

[4] M. De Wachter, M. Matton, K. Demuynck, P. Wambacq, R. Cools, and D. Van Compernolle, "Template based continuous speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 1377–1390, 2007.

[5] J. Gemmeke and B. Cranen, "Noise robust digit recognition using sparse representations," in *Proc. of ISCA 2008 ITRW "Speech Analysis and Processing for knowledge discovery"*, 2008.

[6] J. Gemmeke, L. ten Bosch, L.Boves, and B. Cranen, "Using sparse representations for exemplar based continuous digit recognition," in *EUSIPCO*, Glasgow, Scotland, 2009.

[7] B. Raj and R. M. Stern, "Missing-feature approaches in speech recognition," *Signal Processing Magazine*, vol. 22, no. 5, pp. 101–116, 2005.

[8] M. N. Schmidt and R. K. Olsson, "Linear regression on sparse features for single-channel speech separation," in *Applications of Signal Processing to Audio and Acoustics, IEEE Workshop on (WASPAA)*, 2007.

[9] T. Virtanen, "Spectral covariance in prior distributions of non-negative matrix factorization based speech separation," in *EUSIPCO*, Glasgow, Scotland, 2009.

[10] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, 2003.

[11] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, 2007.

[12] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[13] T. Virtanen, A. T. Cemgil, and S. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *IEEE International Conference on Audio, Speech and Signal Processing*, Las Vegas, USA, 2008.

[14] H.G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. of ISCA ASR2000 Workshop, Paris, France*, 2000, pp. 181–188.

[15] M. Van Segbroeck and H. Van hamme, "Robust speech recognition using missing data techniqies in the prospect domain and fuzzy masks," in *Proc. of IEEE ICASSP*, 2008, pp. 4393–4396.