Label noise robustness and learning speed in a self-learning vocal user interface

Bart Ons, Jort F. Gemmeke and Hugo Van hamme

KULeuven – ESAT, Belgium

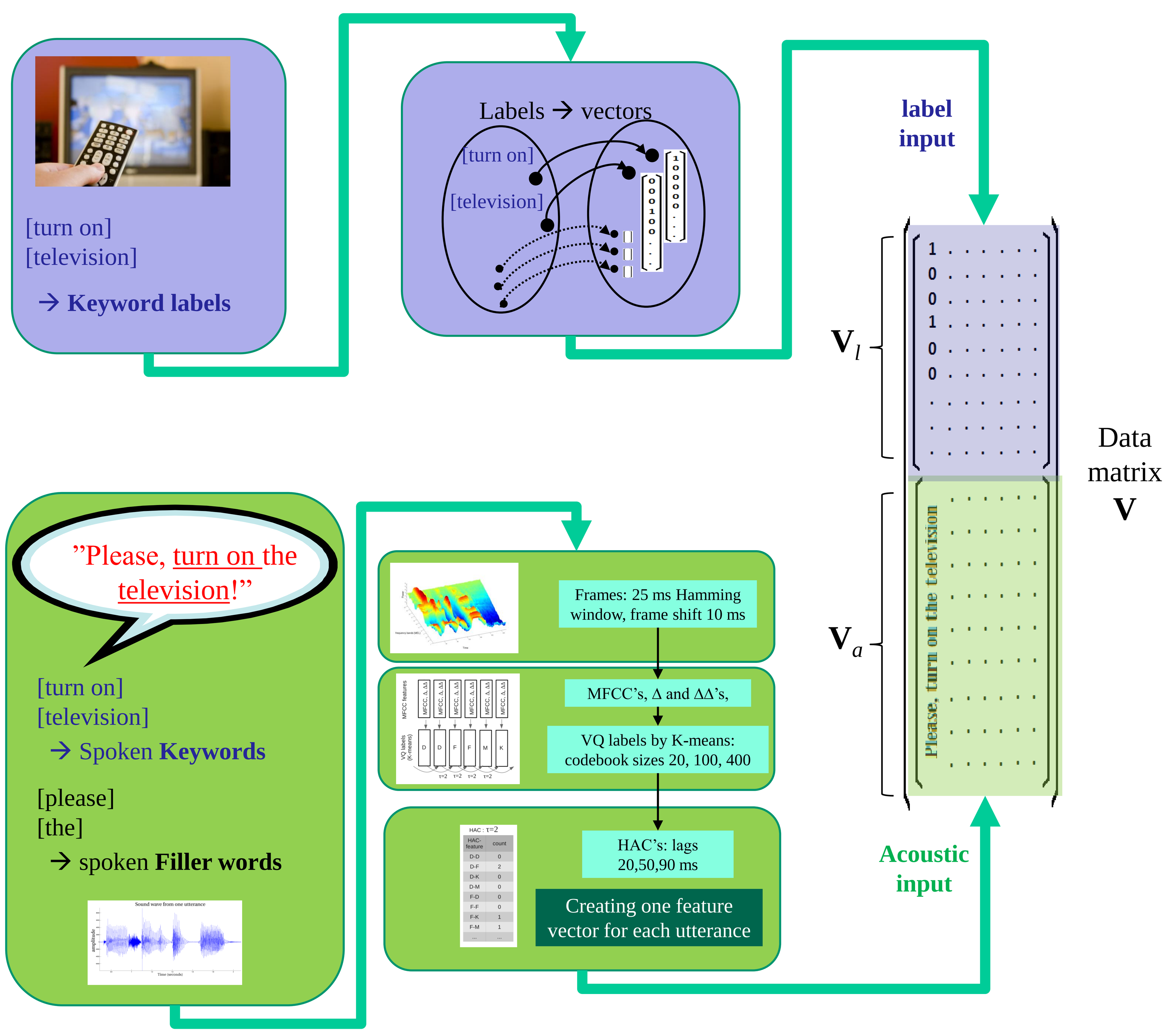
Bart.ons@esat.kuleuven.be

The research was funded by IWT-SBO project ALADIN contract 100049

1. Abstract

- We aim to develop a self-learning vocal user interface that learns to map user-defined spoken commands to demonstrated actions.
- We focus on two requirements:
  - fast learning, i.e. mapping spoken commands on intended actions from a few learning examples
  - Label noise robustness, i.e. limiting the effect of grounding inconsistencies
- We investigated whether supervised non-negative matrix factorization (NMF, see [1, 2]) is able to deal with these requirements.
- We tested keyword spotting for different levels of label noise and training set sizes. Our learning approach is robust against label noise but some improvement regarding fast mapping is desirable.

2. Learning approach of the vocal user interface



3. Formulae

Training

- Aim is to decompose data matrix **V** in the product of two low-dimensional matrices **W** and **H**, with **W** latent representations for keywords and **H** keyword occurrences

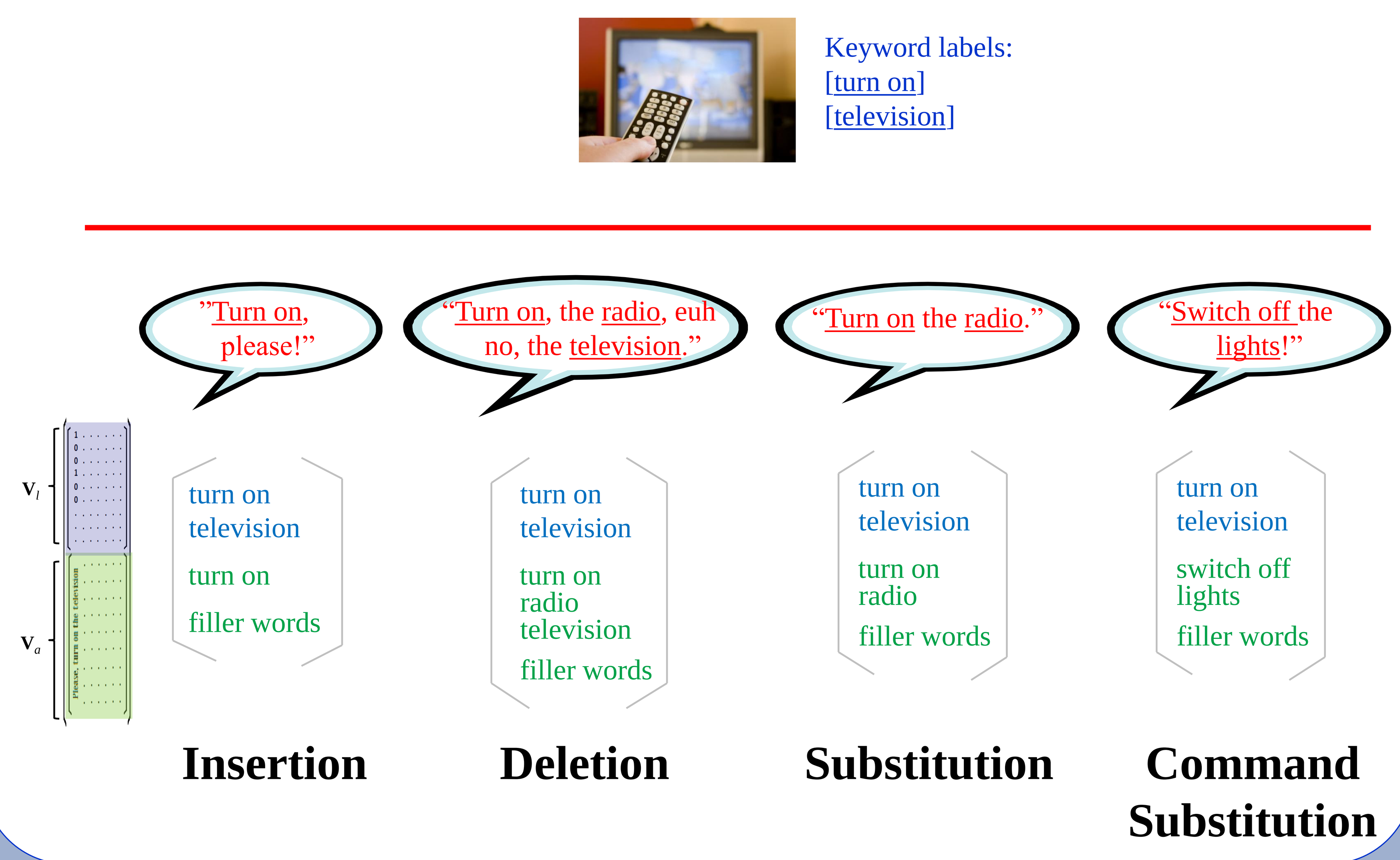
$$\begin{bmatrix} \mathbf{V}_l \\ \mathbf{V}_a \end{bmatrix} \approx \begin{bmatrix} \mathbf{W}_l \\ \mathbf{W}_a \end{bmatrix} \mathbf{H} \qquad \left(\mathbf{H}^*, \mathbf{W}_a^*, \mathbf{W}_l^*\right) = \arg\min_{(\mathbf{H},\mathbf{W}_a,\mathbf{W}_l)} DKL\left(\begin{bmatrix} \mathbf{V}_l \\ \mathbf{V}_a \end{bmatrix} \middle\| \begin{bmatrix} \mathbf{W}_l \\ \mathbf{W}_a \end{bmatrix} \mathbf{H}\right)$$

Testing

- We only have acoustic input $\mathbf{V}_{test,a}$ and aim to find its corresponding label part $\mathbf{V}_{test,l}$, using $\mathbf{W}_a^*$ and $\mathbf{W}_l^*$ from training

$$\mathbf{H}_{test}^* = \arg\min_{\mathbf{H}_{test}} DKL\left(\mathbf{V}_{test,a} \,\|\, \mathbf{W}_a^* \mathbf{H}_{test}\right) \qquad \mathbf{V}_{test,l} = \mathbf{W}_l^* \mathbf{H}_{test}^*$$

4. Label noise



5. Experiments

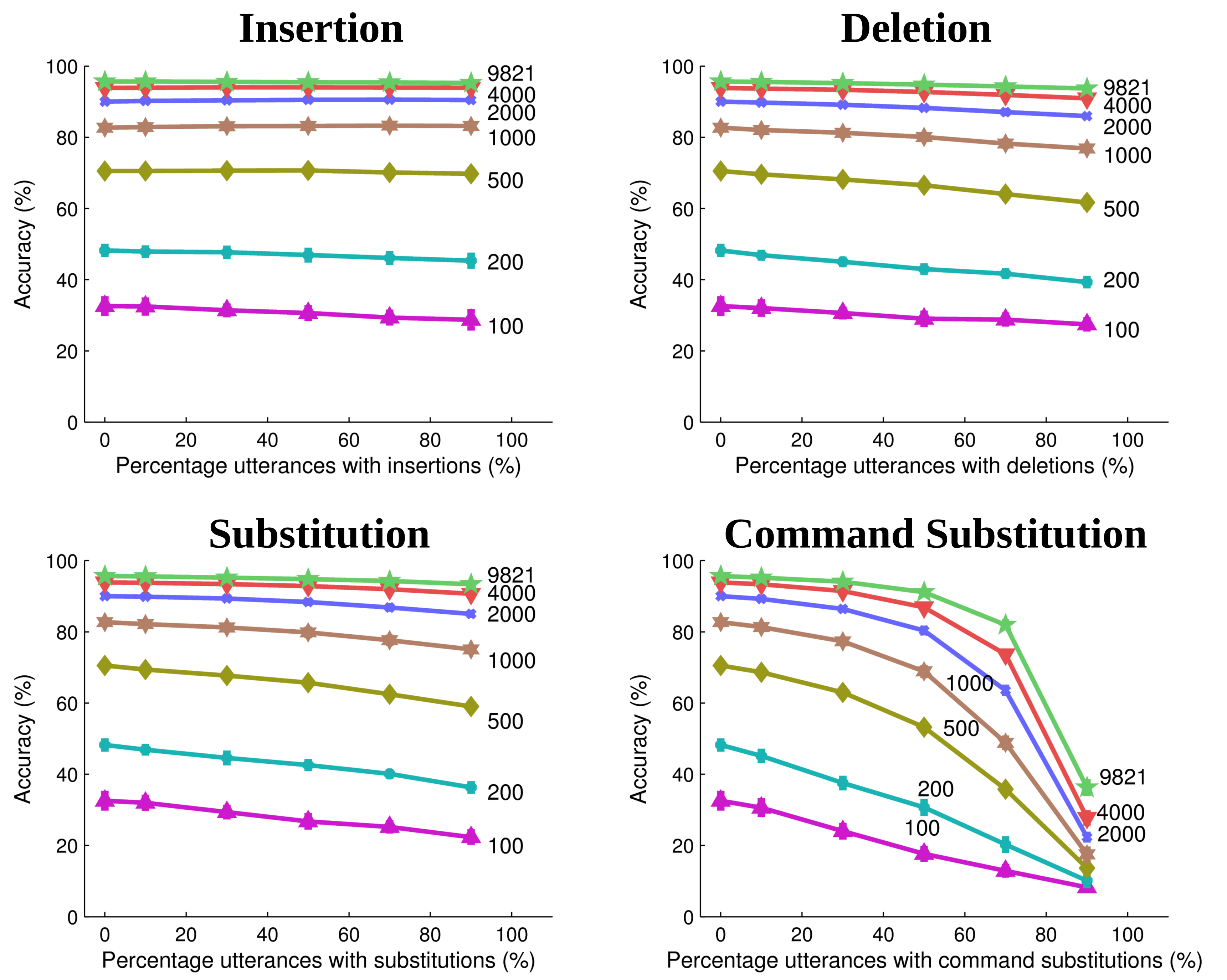
Corpus

- ACORNS English corpus
- Utterances consist of 1 to 4 keywords and filler words
- Vocabulary of 50 keywords

Experimental variables

- Training set sizes: 100, 200, 500, 1000, 2000, 4000 , 9821 utterances
- Label noise : 0, 10, 30, 50, 70 and 90 % of the utterances affected by label noise in the training set
- Four types of label noise, (see box 4)

Results



6. Conclusion

- Improvement regarding learning speed is desirable. In current research we improved the learning speed by using more advanced acoustic input (at present, we are more or less at 80% accuracy for 100 learning examples).
- Leaning is very robust against grounding inconsistencies that take place in the learning environment of the user, allowing more humanized man-machine interactions.

7. References

[1] H. Van hamme, "HAC-models: a novel approach to continuous speech recognition," in *Proc. Interspeech 2008*, Brisbane, Australia, 2008.

[2] J. Driesen, J.F. Gemmeke, and H. Van hamme, "weakly supervised keyword learning using sparse representations of speech," in *Proc. ICASSP*, pp. 5145–5148. Kyoto, Japan (2012)