DIRAC Training - Report

Joseph Keshet
Computer Science and Engineering
The Hebrew University
Jerusalem 91904, Israel
jkeshet@cs.huji.ac.il

September 24, 2006

We proposed an algorithmic research framework for supervised speech recognition problems that builds on recent advances in kernel methods and large margin classifiers. Speech recognition is a complex task, which is composed of several basic tasks such as frame-based phoneme classification, phonetic segmentation, segment-based phoneme classification and word recognition. We addressed each task theoretically, by proposing a suitable discriminative kernel-based algorithm, and then practically by finding a set of informative features and testing on a real benchmark speech dataset. We have already proposed successful approaches for these tasks: frame-based phoneme classification [2], phonetic segmentation [7] and whole sequence phoneme recognition [6].

During the visit we have focused on the task of whole sequence phoneme recognition and its improvements. We extended our previous work [6] to the task of keyword spotting. In this task, we are provided with a speech utterance and a keyword and the goal is to decide whether the keyword is uttered or not, namely, whether the sequence of phonemes was articulated in the given utterance. A report and a paper with some preliminary results are in preparation.

We have also conducted a study on improving the performance of the phoneme recognizer [6] in terms of accuracy. A known method to improve accuracy in HMM-based systems is to use contextual models. Triphone is a representation of each phoneme in a phoneme sequence by three symbols: the phoneme symbol itself, its preceding phoneme symbol and its succeeding phoneme symbol. This representation takes advantage of the sequential context information by making the label space larger. There are two main

challenges in incorporating such a model in a discriminative setup: first, the number of classes grows by the power of three, which becomes non tractable; second, the sparsity of the data makes it difficult to have many (if any at all) examples for each triphone. To overcome these obstacles we proposed two methods for enhancing the monophone system presented in [6] to a triphoneme system. In the first method we just sum over all possible multiphones up to the level of monophone. That is, the monophone, the biphones and the triphone. In the second method we use a special class for each of the triphones and trim (prune) the number of classes by some criterion derived from the sparsity of the data (see for example [3]). This work is still in progress. A report on the method is in preparation. A different way to incorporate contextual information is by using the multi-rasta set of features [5]. Experiment with the multi-rasta feature have not been conducted yet.

Finally, we tried to improve the computational efficiency of the phoneme recognizer [6]. We proposed a method to efficiently compute the Gaussian kernel by Locality sensitive hashing (LSH) [4, 1]. That is, to approximate the sum of the kernels over all the support set in the discriminative inference, by a sum over a subset of the support. This work is still in progress. We also proposed a method to approximate the kernel operator by k-means.

Acknowledgments

We are grateful to David Grangier, whose many of the ideas are part of this work. We are also in debt to Samy Bengio for the helpful discussions, ideas and thoughts.

References

[1] M. Datar, P. Indyk, N. Immorlica, and V. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the Symposium on Computational Geometry*, 2004.

[2] O. Dekel, J. Keshet, and Y. Singer. Large margin hierarchical classification. In *Proceedings of the Twenty-First International Conference on Machine Learning*, 2004.

[3] O. Dekel, S. Shalev-Shwartz, and Y. Singer. The power of selective memory: Self-bounded learning of prediction suffix trees. In *Advances in Neural Information Processing Systems 17*, 2004.

[4] A. Gionis, P. Indyk, and R. Motwani. Similarity search in high dimensions via hashing. In *25th International Conference on Very Large Databases (VLDB)*, 1999.

[5] Hynek Hermansky and Petr Fousek. Multi-resolution rasta filtering for tandem-based asr. In *Proceedings of Interspeech 2005*, 2005.

[6] J. Keshet, S. Shalev-Shwartz, S. Bengio, Y. Singer, and D. Chazan. Discriminative kernel-based phoneme sequence recognition. In *Interspeech*, 2006.

[7] J. Keshet, S. Shalev-Shwartz, Y. Singer, and D. Chazan. Phoneme alignment based on discriminative learning. In *Interspeech*, 2005.