

Two Self-supervised Learning Techniques for Speech Recognition

Damianos Karakos, Haolang Zhou, Puyang Xu, Sanjeev Khudanpur, Andreas G. Andreou
Department of Electrical and Computer Engineering
Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD, USA
{damianos, haolangzhou, puyangxu, khudanpur, andreou}@jhu.edu

I. ACOUSTIC AND LANGUAGE MODELS IN SPEECH RECOGNITION

Most state-of-the-art speech recognition systems use the well-known maximum a posteriori rule

$$\hat{\mathbf{W}} = \arg \max_{\mathbf{W}} P(\mathbf{A}|\mathbf{W})P(\mathbf{W}),$$

for predicting the uttered word sequence \mathbf{W} , given the acoustic information \mathbf{A} . The *acoustic model* is represented by the conditional distribution $P(\mathbf{A}|\mathbf{W})$, while the *language model* is represented by the prior $P(\mathbf{W})$, and the bulk of the research in speech recognition is on training procedures for these two components [1].

Acoustic modeling is usually done in a computationally efficient way, using the maximum likelihood criterion within the parametric family of Gaussian mixtures. The feature space of state-of-the-art systems is typically in the range of hundreds of dimensions—this is the result of concatenating together spectral features from several consecutive speech frames. Dimensionality reduction to a relatively small number of features is then needed to combat sparsity; the problem is further exacerbated by the presence of significant variation in the training data (ambient noise, speaker, age, accent, etc.) which necessitates estimation of thousands (or even millions) of full-covariance Gaussian distributions. Heteroscedastic Linear Discriminant Analysis (HLDA) [2], which is a generalization of Fisher’s Linear Discriminant Analysis, is typically used for this purpose. It is a maximum-likelihood based technique, with a complexity which is only linear with the size of the training data.

Language modeling is the task of assigning a probability distribution to word sequences; the goal is to create distributions that give higher probability to well-formed (grammatical, sensible) word sequences than to non-well-formed ones. Given the inherent sparsity in the data, it is important to employ sophisticated techniques for assigning non-zero probability to all possible sentences, even those that are not encountered in the training data. A large number of *smoothing* techniques [1], [3] have been reported in the literature, but techniques which scale to billions of words of text are usually the most successful and most frequently used. An example of a popular model is the n -gram model, where the probability of a sequence \mathbf{W} can be written using the Markov approxima-

tion $P(\mathbf{W}) = \prod_{i=1}^n P(W_i|W_{i-1}, \dots, W_{i-n+1})$. Furthermore, *maximum entropy* (log-linear) models are also frequently used, as they allow the incorporation of arbitrary features from the immediate or even global context of a word: $P(\mathbf{W}) = Z^{-1} \exp\{\sum_{j=1}^F \lambda_j f_j(\mathbf{W})\}$, where Z is an appropriate scaling constant that depends on the whole vocabulary.

II. SELF-SUPERVISED MODELS

Self-supervised training techniques of acoustic and language models have recently become popular [4], [5], [6], [7], [8], as they often result in significant gains in word recognition accuracy. Here, we focus on two of them: (i) self-supervised HLDA for acoustic modeling, where the unlabeled data are used in addition to the labeled data in the maximum likelihood objective [6]; and (ii) self-supervised discriminative language modeling, where the unlabeled data are used to estimate a set of confusable words over which the language model is further refined [7]. These are summarized below:

- **Self-supervised HLDA [6]:** The goal of HLDA is to estimate a linear transform that preserves as much of the label-dependent information in the transformed space as possible, while maximizing likelihood. Under a Gaussian assumption on the class-conditional distributions, the objective is the maximization of

$$\sum_c N_c \log p(c) + N \log |\Theta| - \sum_c \frac{N_c}{2} \log |\Theta^{(p)} \Sigma_c (\Theta^{(p)})^\top| - \frac{N}{2} \log |\Theta^{(n-p)} \Sigma (\Theta^{(n-p)})^\top| - \frac{nN}{2} \log(2\pi e),$$

where N is the total number of labeled samples, N_c is the number of samples with label c , Θ is the sought-after transform (a $n \times n$ matrix, of which the first p rows $\Theta^{(p)}$ are used in the final model), and Σ_c is the covariance matrix of the Gaussian for label c .

Maximizing the likelihoods of both labeled *and* unlabeled data using the Expectation-Maximization (EM) algorithm [9] results in an objective function similar to above, but with the parameters of the Gaussian distributions estimated using *probabilistic* labels: at each iteration of EM, each unlabeled data sample is assigned a posterior distribution on the labels (instead of the most likely label), and the Gaussian parameters are updated accordingly. Results with

synthetic data and real data from a vowel classification task show the effectiveness of this “soft” version of HLDA [6].

- **Self-supervised discriminative maximum entropy language models [7]:** Instead of just automatically recognizing the unlabeled data and then augmenting the language modeling training text with the automatic transcriptions (see, e.g., [8]) [7] presents a novel technique: the recognition lattices (or *confusion networks*) at the output of the recognizer are used to derive classes of confusable words, that is, words whose posterior probability is within some pre-specified threshold. These are called *cohort sets*, and mainly comprise acoustically confusable words which appear in similar contexts. Since most of the errors in speech recognition result from not being able to distinguish between such words, it makes sense to use a discriminative objective: find features and associated weights (within the exponential family of distributions) such that the *relative likelihood* of the training data, *compared to their competing cohorts*, is as high as possible. I.e., the objective is the maximization of

$$\sum_{i=1}^N \log \frac{\exp\{\sum_{j=1}^N \lambda_j f_j(w_i|h_i)\}}{\sum_{w' \in C(w_i)} \exp\{\sum_{j=1}^N \lambda_j f_j(w'|h_i)\}},$$

where h_i is the “history” (context) of the i -th word in the training data and $C(w_i)$ is its cohort set. Experiments with a real large vocabulary speech recognition task show a significant improvement over a strong 4-gram Kneser-Ney language model [7].

ACKNOWLEDGMENT

We thank Scott Novotney for his pointers to recent self-supervised speech recognition literature.

REFERENCES

- [1] F. Jelinek, *Statistical Methods for Speech Recognition*, MIT Press, 1997.
- [2] N. Kumar and A. G. Andreou, “Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition,” *Speech Communication*, vol. 26, pp. 283–297, 1998.
- [3] S. F. Chen and J. Goodman, “An empirical study of smoothing techniques for language modeling,” in *Proceedings of the 34th Annual Meeting of the ACL*, 1996, pp. 310–318.
- [4] L. Lamel, J. luc Gauvain, and G. Adda, “Lightly supervised and unsupervised acoustic model training,” *Computer Speech and Language*, vol. 16, pp. 115129, 2002.
- [5] J. Ma, S. Matsoukas, O. Kimball, and R. Schwartz, “Unsupervised training on large amounts of broadcast news data,” in *Proceedings of the 2006 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-06)*, Toulouse, France, May 2006.
- [6] H. Zhou, D. Karakos, and A. G. Andreou, “A semi-supervised version of heteroscedastic linear discriminant analysis,” in *Proceedings of the 2009 Conference of the International Speech Communication Association (Interspeech-09)*, Brighton, UK, September 2009.
- [7] P. Xu, D. Karakos, and S. Khudanpur, “Self-supervised discriminative training of statistical language models,” in *Proceedings of the 2009 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU-09)*, Merano, Italy, December 2009.
- [8] S. Novotney, R. Schwartz, and J. Ma, “Unsupervised acoustic and language model training with small amounts of labeled data,” in *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-09)*, Taipei, Taiwan, April 2009.
- [9] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 39, pp. 1–38, 1977.