Anti-Models: An Alternative Way to Discriminative Training

Jan Vaněk* and Josef Psutka

University of West Bohemia in Pilsen, Univerzitní 22, 306 14 Pilsen
Faculty of Applied Sciences, Department of Cybernetics
vanekyj@kky.zcu.cz

Abstract. Traditional discriminative training methods modify Hidden Markov Model (HMM) parameters obtained via a Maximum Likelihood (ML) criterion based estimator. In this paper, anti-models are introduced instead. The anti-models are used in tandem with ML models to incorporate a discriminative information from training data set and modify the HMM output likelihood in a discriminative way. Traditional discriminative training methods are prone to over-fitting and require an extra stabilization. Also, convergence is not ensured and usually "a proper" number of iterations is done. In the proposed anti-models concept, two parts, positive model and anti-model, are trained via ML criterion. Therefore, the convergence and the stability are ensured.

Keywords: ASR, HMM, Acoustic Modeling, Discriminative Training, Anti-Models, MMI, MCE, MPE

1 Introduction

Discriminative training (DT) techniques have been shown to outperform the ML-based training in automatic speech recognition (ASR). But they require a proper tuning and use a number of heuristics [3]. Moreover, they usually do not converge and maximization of a training criterion may not lead to maximum recognition accuracy with unseen data. ASR systems used acoustic models with a reduced complexity in past due to limited computing power. DT techniques gain is better with the less-complex models trained from the same amount of training data [7]. The more-complex models are more sensitive to a DT setup and require a finer tuning to get a significant gain over the ML models. Nowadays, a multi-core computer architecture has enough computing power to run any single diagonal-covariance acoustic model in real-time and, with a GPU acceleration, even multiple models [1] or full-covariance models [2]. Therefore, the more and more-complex HMMs (even with full-covariance matrices) are used and it is obvious that some simpler and more robust DT technique could be helpful.

In this paper, a concept of anti-models is introduced. The idea is simple: Instead of modification of the Gaussian parameters to fit the model into an unnatural (non-Gaussian) shape using negative statistic or gradient methods, we directly construct the

* This research was supported by the Grant Agency of the Czech Republic, project No. GAČR GBP103/12/G084.

anti-model, model of the data that belongs to the others HMM states. The idea is general and can be adopted to all major DT criteria, e.g. Maximal Mutual Information (MMI), boosted-MMI, Minimum Classification Error (MCE), and Minimum Phone Error (MPE). To keep a clarity and to fit into limited paper size, the MMI case is shown in detail in this paper. Nevertheless, the derivation for the other criteria is analogous.

This paper is organized as follows. A brief overview of DT techniques is given in Section 2. The concept of anti-models is introduced in Section 3. The derivation of the anti-models with the MMI criterion is described in Section 4. Experiments and results are presented and discussed in Section 5.

2 Dicsriminative Training

In principle, the ML based training is a machine learning method from positive examples only. In contrast to ML, discriminative approaches take into account an information about class competition during the training. This extra information may improve results, but it brings an extra computation burden also. A short review of the most frequently used discriminative criteria follows.

2.1 Maximum Mutual Information - MMI

In the MMI case, a training algorithm seeks to maximize the posterior probability of the correct utterance given the used models [8]:

$$\mathcal{F}_{MMI}(\lambda) = \sum_{r=1}^{R} \log \frac{P_\lambda(O_r|s_r)^\kappa P(s_r)^\kappa}{\sum_S P_\lambda(O_r|s)^\kappa P(s)^\kappa}, \tag{1}$$

where λ represents the acoustic model parameters, O_r is the training utterance feature set, s_r is the correct transcription for the r'th utterance, κ is the acoustic scale which is used to amply confusions and herewith increases the test-set performance. $P(s)$ is a language model part. Optimization of the criterion (1) requires to generate lattices or many-hypotheses recognition run with appropriate language model. The lattices generation is highly time consuming. Furthermore, these methods require good correspondence between training and testing dictionary and language model. If the correspondence is weak, e.g. there are many words which are only in the test dictionary then the results of these methods are not good. In this case, we can employ Frame-Discriminative training (MMI-FD), which is independent on the used dictionary and language model [6]. In addition, this approach is much faster.

Optimization of the MMI objective function uses Extended Baum-Welch update equations [5] and it requires two sets of statistics. The first set, corresponding to the numerator (num) of the equation (1), is the correct transcription. The second one corresponds to the denominator (den) and it is a recognition/lattice model containing all possible words. An accumulation of statistics is done by the forward-backward algorithm on reference transcriptions (numerator) as well as generated lattices (denominator). The Gaussian means and variances are updated as follows [6]:

$$\hat{\mu}_{jm} = \frac{\Theta_{jm}^{num}(O) - \Theta_{jm}^{den}(O) + D_{jm}\mu'_{jm}}{\gamma_{jm}^{num} - \gamma_{jm}^{den} + D_{jm}} \tag{2}$$

$$\hat{\sigma}^2_{jm} = \frac{\Theta^{num}_{jm}(O^2) - \Theta^{den}_{jm}(O^2) + D_{jm}(\sigma'^2_{jm} + \mu'^2_{jm})}{\gamma^{num}_{jm} - \gamma^{den}_{jm} + D_{jm}} - \mu^2_{jm}, \tag{3}$$

where j and m are HMM-state and Gaussian indexes, respectively, γ_{jm} is the accumulated occupancy of the Gaussian, $\Theta_{jm}(O)$ and $\Theta_{jm}(O^2)$ are a posterior probability weighted by the first and the second order accumulated statistics, respectively. Gaussian-specific stabilization constants D_{jm} are set to maximum of (i) double of the smallest value which ensures positive estimated variances, and (ii) value $E\gamma^{den}_{jm}$, where constant E determines the stability/learning-rate and it is a compromise between stability and number of iteration which is needed for well-trained models [9]. To bring a stability to HMM states with low-data, I-smoothing was introduced in [10] that uses a prior from the ML model or the MMI model from the previous iteration.

2.2 Boosted-MMI

The Boosted-MMI [14] is a modification of the MMI method. The denominator lattice trajectories are weighted by an error in this method. The error is defined in the same way like in MPE or MWE (see below). Therefore, the Boosted-MMI is a combination between MMI and MPE/MWE.

2.3 Minimum Classification Error - MCE

Another popular criterion is MCE [11–13]. The MCE criterion directly minimizes an error in a recognized word sequence. In contrast, MMI maximizes a probability of the correct sequence against others. The non-smooth classification error is smoothed via a sigmoid function to allow gradient-based optimization. It operates with n-best lists or with lattices that are more suitable for large vocabulary continuous speech recognition.

2.4 Minimum Word Error - MWE

The MWE criterion is defined formally similar to (1), but with a word error incorporation:

$$\mathcal{F}_{MWE}(\lambda) = \sum_{r=1}^{R} \log \frac{\sum_S P_\lambda(O_r|s_r)^\kappa P(s_r)^\kappa \text{RawAccuracy}(s)}{\sum_S P_\lambda(O_r|s)^\kappa P(s)^\kappa}, \tag{4}$$

where RawAccuracy(s) is the error rate of the word sequence s, κ is the acoustic model weight. This criterion is a weighted mean of the correct recognized words out of all other possible word sequences. Maximization of the MWE criterion improve a number of correctly recognized words in the most probable word sequences.

2.5 Minimum Phone Error - MPE

The MWE shows poor ability to generalize the training performance to unseen data. Therefore, more robust phone-based criteria were introduced. The maximization is aimed to the phone-level here [7]. Formal description of the MPE criterion is identical to (4). Only the error part RawAccuracy(s) is different in MPE. It express a relative count of correct recognized phones. An approximation to the error is used in practice for efficiency reasons [15].

2.6 Minimum Phone Frame Error - MPFE

Another approximation to the phone error was introduced in [16]. The MPE described above does not penalize deletion errors sufficiently. In addition, dynamic range of phone RawAccuracy(s) is typically quite narrow, which makes MPE occupancies considerably lower than MMI occupancies [10]. This may lead to an MPE robustness problem when training data are not abundant. MPFE uses phone-lattices and modified criterion to overcome the MPE shortcomings.

3 Concept of Anti-Models

The traditional DT techniques have two main shortcomings: Low stability and proneness to over-fitting. An implementation of the DT training that successfully copes with the shortcomings often seems to be more art than science. In contrast, the ML training is stable and there is a considerable experience how to prevent the over-fitting [17–22]. The concept of the anti-models is based on advantages of the ML training and in addition of the HMM state-concurrency information. Other-states data that have high likelihood in the model of the particular state are modeled separately by the anti-model. Thus, each state has standard ML model and the anti-model. The final state observation probability $p(o_t)$ of the feature vector $p(o_t)$ is calculated as follows

$$p(o_t) = \frac{p_{ML}(o_t)^2}{p_{ML}(o_t) + w_A p_A(o_t)}, \tag{5}$$

where $p_{ML}(o_t)$ is the output probability of the ML model and $p_A(o_t)$ is probability of the anti-model. w_A is the weight of the anti-model that is proportional to an amount of the wrongly modeled data. An illustration of the anti-model concept is in Figure 1. The figure shows that the anti-model modifies only the part of the distribution, where the wrongly modeled data were observed. In contrast, MMI modifies the entire distribution even in the part where no data were observed and therefore, it may be less optimal for unseen data. The figure also shows that the anti-model concept is able to model even non-Gaussian distributions. A log-domain is usually used in practice for HMM observation probabilities evaluation. The log form of the equation (5) is

$$\log(p(o_t)) = 2\log(p_{ML}(o_t)) - \log(p_{ML}(o_t) + w_A p_A(o_t)). \tag{6}$$

The sum of two probabilities evaluated in the log-domain can be calculated in a more robust and faster way:

$$\begin{aligned} \log(e^{lp_1} + e^{lp_2}) &= \max(lp_1, lp_2) + \log(1 + e^{(\min(lp_1,lp_2) - \max(lp_1,lp_2))}) \\ &= \max(lp_1, lp_2) + f(\min(lp_1, lp_2) - \max(lp_1, lp_2)), \end{aligned} \tag{7}$$

where lp_1 and lp_1 are two probabilities evaluated in the log-domain and the function $f(\min(lp_1, lp_2) - \max(lp_1, lp_2))$ is a smooth function that fast limits to zero for larger values. The function can be effectively approximated via Taylor series decomposition or calculated via lower-accuracy hardware implemented GPU instructions. Note, that

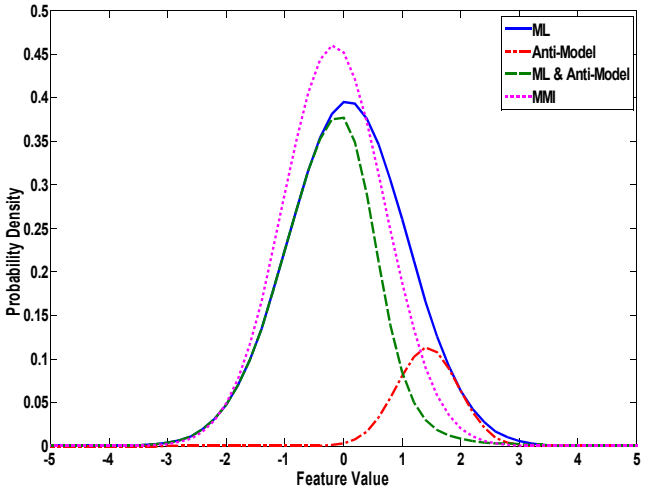

Fig. 1. Illustration of the anti-model concept. Probability density functions are compared for ML, MMI, anti-model, and ML & anti-model as the application of equation 6.

this approach used to be implemented in the HMM evaluation already, thus it is simple to reuse it for the anti-models also. The resulting anti-models based HMM has about twice as many parameters to estimate. However, the more reliable ML training is used to the estimation. Note that the ML model and the anti-model of the state do not need to have the equal number of components and a proper model complexity can be chosen for both parts.

4 MMI-Based Anti-Models

Derivation of the anti-model parameters estimation for MMI is straight forward and simple. The equations are the same like for ML training. Only the denominator statistics of the MMI criterion is used:

$$\hat{\mu}_{Ajm} = \frac{\Theta^{den}_{jm}(O)}{\gamma^{den}_{jm}} \tag{8}$$

$$\hat{\sigma}^2_{Ajm} = \frac{\Theta^{den}_{jm}(O^2)}{\gamma^{den}_{jm}} - \mu^2_{jm}. \tag{9}$$

No additional stabilization nor the I-smoothing is needed. There is only one difference, the denominator statistic for the anti-model does not contain the state's own data - the positive examples, only the negative ones. Estimation of the anti-model weight w_{Ajm} is also simple

$$\hat{w}_{Ajm} = \frac{\gamma^{den}_{jm}}{\gamma^{num}_{jm}}. \tag{10}$$

5 Experiments

A comparison of the proposed concept of the anti-models with the traditional ML and MMI criteria was done on a simple ASR task.

5.1 Speech Data, Processing, and Test Description

A part of UWB_S01 corpus [4] was used for experiments purposes. Data from first 100 speakers (57 males, 43 females) were used as a training part. Another 100 speakers (64 males, 36 females) make a test part. The digitization of an analogue signal was provided at 22.05 kHz sample rate and 16-bit resolution format. In order to extract features, Mel-frequency cepstral coefficients (MFCCs) were utilized, 15 dimensional feature vectors were extracted each 10 ms utilizing a 32 ms hamming window, including the energy coefficient. Then, Cepstral Mean Normalization (CMN) was applied, and Δ, Δ^2 coefficients were added. A 3 state HMM based on triphones with 425 states in total and 8 component GMM with diagonal covariances in each of the states was trained via ML criterion. In the case of anti-models, 8 component GMMs were trained for both the parts, ML and anti-model. The anti-models as well as the MMI models are based on MMI-FD criterion. In the case of MMI, three variants that differ in setup of the stabilization constant E were trained (see 2.1). Two variants with the fixed E equal 1 and 2 were done and one variant with dynamic E equal $\sqrt{\text{iteration}}$ has been added. I-smoothing was employed for all the MMI variants with $\tau = 100$.

To test the performance a simple 476-words zero-gram language model (LM) with no OOV was used. The simple LM was used to boost sensitivity to the acoustic part of the system. The zero-gram variant of the recognizer published in [23] was used and Word Error Rate (WER) was evaluated.

5.2 Results

The proposed concept of the anti-models were compared to the ML baseline and the MMI discriminative training. Both, MMI and anti-models, was initialized by the ML baseline model. The results are shown in Figure 2. It is clear that MMI as well as the anti-models outperform ML training. MMI produces very good models just after a single iteration on this task. However, the stability is a problem and after a few additional iterations WER increases. In contrast, the anti-models needs at least two iterations to proper results, but it keeps the achieved level.

6 Conclusion

The alternative way to discriminative training, the concept of the anti-models, was introduced in this paper. Advantages and disadvantages were discussed. The main advantage is the stability and robustness that is brought by ML training. The main disadvantage is a more complex final HMM that requires a higher computation power to evaluate. The concept was tested on the ASR experiment. The results confirm equivalent WERs but the better stability in comparison to MMI.

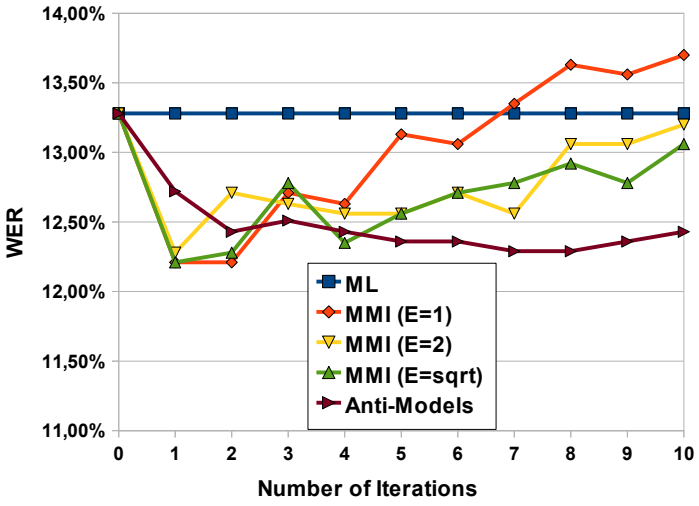

Fig. 2. Results of the ASR experiment. WER was evaluated for ML, three variants of MMI, and for the anti-models.

References

1. Vanek J., Trmal, J., Psutka, J.V., Psutka, J.: Optimized Acoustic Likelihoods Computation for NVIDIA and ATI/AMD Graphics Processors. In: IEEE Transactions on Audio, Speech and Language Processing, Vol. 20, 6, pp. 1818-1828, 2012.
2. Vanek, J., Trmal, J., Psutka, JV., Psutka, J.: Full Covariance Gaussian Mixture Models Evaluation on GPU. In: Proc. IEEE ISSPIT, Vietnam, Ho Chi Minh City, 2012.
3. Heigold G., Schl uter, R., Ney, H., Wiesler, S.: Discriminative Training for Automatic Speech Recognition: Modeling, Criteria, Optimization, Implementation, and Performance. In: IEEE Signal Processing Magazine, 11/2012, Vol. 29, pp. 58-69. DOI:10.1109/MSP.2012.2197232.
4. Radová, V., Psutka, J.: UWB_S01 Corpus - A Czech Read-Speech Corpus. In: Proc. of the ICSLP 2000, pp. 732–735, Beijing, China, 2000.
5. Normandin, Y., Morgera, D.: An Improved MMIE Training Algorithm for Speaker-Independent Small Vocabulary, Continuous Speech Recognition. In: Proc. of the IEEE, ICASSP 1991, pp. 537-540. Toronto, Canada, 1991.
6. Kapadia, S.: Discriminative Training of Hidden Markov Models. In: PhD thesis, Cambridge University, Department of Engineering, 1998.
7. Povey, D.: Discriminative Training for Large Vocabulary Speech Recognition. In: PhD thesis, Cambridge University, Department of Engineering, 2003.
8. Bahl, L.R. at al.: Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition. In: ICASSP 1986, Tokyo, Japan, 1986.
9. Povey, D. at al.: Improved discriminative training techniques for large vocabulary continuous speech recognition. In: ICASSP 2001, Salt Lake City, Utah, USA, 2001.
10. Povey, D., Woodland, P.C.: Minimum Phone Error and I-Smoothing for Improved Discriminative Training. In: Proc. ICASSP 2002, Orlando, USA, 2002.
11. Katagiri, S., Lee, C.-H., Juang, B.-H.: New Discriminative Training Algorithms Based on the Generalized Descent Method. In: Proc. IEEE Neural Networks for Signal Processing, pp. 299-308, 1991.

12. Juang, B.-H., Katagiri, S.: Discriminative Learning for Minimum Error Classification. In: IEEE Transactions on Acoustic, Speech and Signal Processing, Vol. 40, no. 12, pp. 3043-3054, 1992.
13. Chou, W., Juang, B.-H., Lee, C.-H.: Segmental GDP Training of HMM Based Speech Recognizer. In: Proc. ICASSP, vol. 1, pp. 473-476, 1992.
14. Povey, D., Kanevsky, D., Kingsbury, B., Ramabhadran B., Saon, G., Visweswariah, K.: Boosted MMI for Model and Feature-Space Discriminative Training. In: Proc. ICASSP 2008, Las Vegas, USA, 2008.
15. Povey, D., Woodland, P.C., Gales, M.J.F.: Discriminative MAP for Acoustic Model Adaptation. In: Proc. ICASSP 2003, Hong Kong, 2003.
16. Zheng, J., Stolcke, A.: Improved Discriminative Training. Using Phone Lattices. In: Proc. Eurospeech 2005, Lisbon, Portugal 2005.
17. Bell, P.: Full Covariance Modelling for Speech Recognition. In: Ph.D. Thesis, The University of Edinburgh.
18. Lee, Y., Lee, K.Y., Lee, J.: The Estimating Optimal Number of Gaussian Mixtures Based on Incremental k-means for Speaker Identification. In: International Journal of Information Technology, Vol.12, No.7, pp. 13-21, 2006.
19. Figueiredo, M., Leitão, J., Jain, A.: On Fitting Mixture Models. In: Proc. EMMCVPR 1999, pp. 54-69, Lecture Notes In Computer Science, Springer-Verlag London.
20. Mclachlan, G.J., Peel, D.: On a Resampling Approach to Choosing the Number of Components in Normal Mixture Models. In: Computing Science and Statistics, Vol. 28, pp. 260-266, 1997.
21. Paclík, P., Novovičová, J.: Number of Components and Initialization in Gaussian Mixture Model for Pattern Recognition. In: Proc. Artificial Neural Nets and Genetic Algorithms, pp. 406-409, Springer-Verlag Wien, 2001.
22. Vanek, J., Machlica, L., Psutka, JV., Psutka, J.: Covariance Matrix Enhancement Approach to Train Robust Gaussian Mixture Models of Speech Data In: Proc. Speech and Computer 2013, Lecture Notes in Computer Science, vol. 8113, p. 92-99, Springer, 2013.
23. Pražák, A., Psutka, J., Hoidekr, J., et al.: Automatic online subtitling of the Czech parliament meetings, Lecture Notes in Artificial Intelligence, pp. 501–508 (2006).