DISCRIMINATIVELY TRAINED SPOKEN DOCUMENT SIMILARITY MODELS AND THEIR APPLICATION TO PROBABILISTIC LATENT SEMANTIC ANALYSIS

K. Thambiratnam, F. Seide and P. Yu

Microsoft Research Asia, 5F Beijing Sigma Center, No. 49 Zhi Chun Rd., 100080 Beijing, P.R.C.

ABSTRACT

This paper presents a novel framework for discriminatively training spoken document similarity models. Traditional similarity methods such as Vector Space Modeling and Probabilistic Latent Semantic Analysis suffer from a mismatch in modeling and evaluation objective functions. This work proposes reconciling this mismatch by using a discriminative training process in conjunction with prior knowledge of known document relationships to train an ensemble of spoken document similarity models. The reported experiments demonstrate dramatic improvements in mAP performance for the tasks of related document search and query-by-document retrieval, and highlight the ability of the resulting models to better generalize to unseen topics and unseen documents.

Index Terms— Document Similarity, Spoken Document Retrieval, Information retrieval, Modeling, Pattern Recognition

1. INTRODUCTION

The continuing growth in the volume and the importance of multimedia content demands the development of more robust methods for searching and exploring large multimedia databases. Though traditionally, large multimedia corpora have been accessed through query search interfaces, there is a growing interest in more passive retrieval modes that use *implicit* queries, such as automatic content recommendation systems and related-content browsers. Passive retrieval modes are particularly important in the spoken/multimedia document domain, since typically such media is consumed in a more passive fashion than text. Commercial examples of these include related video listings on online multimedia portals and TiVo content recommendations. Content-based contextual advertising is another example, where implicit queries based on the current multimedia document are used to select related advertising.

Document-to-document similarity is typically the technology of choice for such applications. The goal here is to use an entire spoken document as a query for retrieving similar or relevant documents from a corpus. Typical approaches use automatic speech recognition to first generate text transcriptions, and then apply text-domain techniques, such as Vector Space Modeling (VSM) [1], Latent Semantic Analysis (LSA) [2] or Probabilistic Latent Semantic Analysis (PLSA) [3], to compute document similarity.

A fundamental issue in many document similarity techniques is the mismatch between modeling and evaluation objectives. For example, both LSA and PLSA model a document-word co-occurrence matrix, however this criteria is not directly related to the document similarity task. VSM is a heuristic approach with little scientific justification for computing similarity. Mismatch between modeling and evaluation objectives leads to sub-optimality, and reconciling these differences has been shown to be beneficial for a variety of tasks, including hand-writing recognition [4] and object recognition [5]. Two speech domain examples are Minimum Classification Error (MCE) [6] and Minimum Phone Error (MPE) [7] training, well accepted techniques for improving acoustic models for speech recognition. Using a training objective function that is consistent with the evaluation task metrics allows direct optimization of a model for the task at hand, and thus intuitively results in better performing systems.

Thus, this work proposes a generalized discriminative training framework for reconciling the training and evaluation objectives for document similarity. Prior information about document relations and more importantly *non*-relations, is used to discriminatively train an ensemble of document similarity classification models. The aim is to use this supervised information source to train the modelset to not only learn how to better classify related documents, but additionally, to reduce competition with classifiers of unrelated documents.

The framework is then applied to the PLSA model structure. PLSA is applied here because it is felt that for *spoken* document retrieval, the PLSA co-occurrence modeling provides a layer of robustness to speech recognition errors. It should be noted though that the proposed framework can be applied to other model structures.

The paper is organized as follows. A brief review of related theory is first provided in section 2, followed by details of the proposed technique in section 3. Evaluation results are then reported in section 4 before the paper concludes in section 5.

2. BACKGROUND

Of the many document similarity techniques proposed in the literature, VSM [1] is arguably the most common. In VSM, each document is represented by a vector $\mathbf{x^i}$, where $x_k^i = TF(d_i, w_k) \times \sqrt{IDF(w_k)}$ is commonly used to represent the relative frequency of word w_k in document d_i. Here $TF(d_i, w_k) = P(w_k|d_i) = n(d_i, w_k)/\sum_{k'=1}^{K} n(d_i, w_{k'})$, called the Term Frequency, is the intra-document word frequency and $IDF(w_k) = \log D/N_D(w_k)$ is the well known Inverse Document Frequency (IDF) global term weighting. $n(d_i, w_k)$ is the number of occurrences of w_k in document d_i, D is the number of database documents, and $N_D(w_k)$ is the number of documents in the database in which word w_k occurs at least once. Document similarity can then be computed using the cosine distance measure, SIM_{VSM}:

$$SIM_{VSM}(\mathbf{x^1}, \mathbf{x^2}) = \frac{\mathbf{x^1} \cdot \mathbf{x^2}}{|\mathbf{x^1}|\,|\mathbf{x^2}|} \qquad (1)$$

2.1. Probabilistic Latent Semantic Analysis

Probabilistic Latent Semantic Analysis (PLSA), proposed by [3], is a probabilistic approach to Latent Semantic Analysis (LSA) [2]. Both LSA and PLSA derive latent semantic factors that allow documents to be conceptualized in a high-level semantic space, however PLSA uses a more statistically sound foundation. PLSA attempts to discover and model patterns in word co-occurrence by modeling the document-word co-occurrence matrix using the joint distribution, $P(d, w)$, and the latent semantic variable, z. Assuming conditional independence of d and w with respect to z, this gives

$$P(d_i, w_k) = \sum_{j=1}^{J} P(z_j)P(d_i|z_j)P(w|z_j) \tag{2}$$

The individual PDFs, $P(z)$, $P(d|z)$ and $P(w|z)$ are trained using Maximum Likelihood (ML) techniques such as Expectation Maximization (EM). The resulting PDFs can then be used to compute a number of measures, including the factor-space representation of each document $\mathbf{d^i}$, $d^i_j = p(z_j|d_i)$. Note the similarity here with the TF document vector, since TF can be written as $P(w_k|d_i)$. Thus, $\mathbf{d^i}$ is referred to as the Expected Factor Frequency (EFF) document vector within this work, where *expected* emphasizes the fact that $\mathbf{d^i}$ is only a probabilistic estimate of the true Factor Frequency vector.

PLSA similarity can then be computed using a VSM-inspired approach [3] by computing the cosine distance of the EFF vectors:

$$SIM_{PVSM}(\mathbf{d^1}, \mathbf{d^2}) = \frac{\mathbf{d^1} \cdot \mathbf{d^2}}{|\mathbf{d^1}|\,|\mathbf{d^2}|} \tag{3}$$

The above similarity measure requires an EFF representation for unseen query documents in order to compute similarity with documents within a database. Typically, a query document, q, is approximated or *folded* into the factor space using the PLSA PDFs. One approach is to fix $P(w|z)$ and $P(z)$ and to then use EM to estimate $P(z|q)$ [3]. Here, the empirical distribution $\tilde{P}(q, w_k)$ is approximated using the query word counts $n(q, w_k)$. Alternately, the empirical word distribution $\tilde{P}(w_k|q)$ derived from the query TF vector $\mathbf{y}$, $y_k = p(w_k|q) = TF(q, w_k)$ can be used to probabilistically predict a query's representation, using $P(z_j|q) = \sum_{k=1}^{K} P(z_j|w_k)\tilde{P}(w_k|q)$ assuming that $P(z_j|w_k, q) \approx P(z_j|w_k)$.

3. DISCRIMINATIVE DOCUMENT SIMILARITY MODELS

This section presents a generalized discriminative training framework for training an ensemble of document models using a training objective that is consistent with the document similarity evaluation task. This is done by using prior knowledge about document relationships to train the ensemble of document models to minimize classification error given this relationship information.

Let $\mathbf{\Lambda} = \left[\lambda^{\mathbf{i}}\right]$ be an ensemble of target document models representing each document, d_i in a database of D documents. Each model tries to capture the multitude of topics and other bits of information that are useful for determining document similarity. Document similarity thus attempts to find the most similar documents to q using the modelset $\mathbf{\Lambda}$. If some knowledge of inter-document relationships is known during the training of $\mathbf{\Lambda}$ then it is hoped that the proposed training algorithm can adjust each model, $\lambda^{\mathbf{i}}$, for each training document, $\mathbf{x^n}$, in such a way that:

1. Classification for $\mathbf{x^n}$ is improved for $\lambda^{\mathbf{i}}$ if it is related to $\mathbf{x_n}$.
2. More importantly, competitiveness with other classes that are related to $\mathbf{x_n}$ is reduced, if $\lambda^{\mathbf{i}}$ is *NOT* related to $\mathbf{x_n}$.

The intention here is to not only train individual document models, $\lambda^{\mathbf{i}}$ to better classify related query documents, but *more importantly*, to *NOT* compete against the classifiers of unrelated documents. For example, if d_1 is known *NOT* to be related to d_2, then training $\lambda^{\mathbf{1}}$ to *NOT* compete with $\lambda^{\mathbf{2}}$ will hopefully result in less competition for other unseen query documents related to d_2.

The training process is formulated using the Generalized Probabilistic Descent (GPD) framework. The basic theory of GPD is not covered here since a thorough review is available in [8]. GPD training is applied here to minimize the expected number of errors (ie. the expected loss) for the ensemble of document models, $\mathbf{\Lambda} = \left\{\lambda^{\mathbf{1}}, \ldots, \lambda^{\mathbf{A}}\right\}$, given a set of training samples, $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_N}\}$. Expected loss is defined here in terms of a set of individual class loss functions, $l_a(\mathbf{x_n})$, that measure the loss of classifying query $\mathbf{x_n}$ as class C_a, where class C_a is modeled by document model $\lambda^{\mathbf{a}}$. Since $\mathbf{x_n}$ can be related to multiple document classes, the expected loss is defined using the per-sample average loss as follows:

$$L(\mathbf{\Lambda}) = \frac{1}{N} \sum_{n=1}^{N} \frac{\sum_{a=1}^{A} l_a(\mathbf{x_n})\delta(\mathbf{x_n} \in C_a)}{\sum_{a=1}^{A} \delta(\mathbf{x_n} \in C_a)} \tag{4}$$

Here the indicator function $\delta(cond) = 1$ when $cond$ is true and 0 otherwise. Averaging is performed on a per-training sample basis across all document classes that are related to a training sample. This allows multiple relationships to be considered per training sample while preventing the expected loss function from being dominated by training samples with many related classes.

The class loss function is defined using a smoothed step function parameterized on the class discrimination function, $d_k(\mathbf{x_n})$, which in turn is derived from the class distance function $g_a(\mathbf{x_n})$.

$$l_a(\mathbf{x_n}) \quad = \quad \frac{1}{1 + e^{-(\alpha d_a(\mathbf{x_n}) + \beta)}} \tag{5}$$

$$d_a(\mathbf{x_n}) \quad = \quad g_a(\mathbf{x_n}) - \left[\frac{\sum_{b=1}^{A} \left[g_b(\mathbf{x_n})\delta(\mathbf{x_n} \notin C_b)\right]^{-\mu}}{\sum_{b=1}^{A} \delta(\mathbf{x_n} \notin C_b)}\right]^{-\frac{1}{\mu}} \tag{6}$$

$$g_a(\mathbf{x_n}) \quad = \quad e^{-\gamma SIM(\lambda^{\mathbf{a}}, \mathbf{x_n})} \tag{7}$$

The functional form used for $d_k(\mathbf{x_n})$ provides a smoothed $min()$ function, where smoothness is controlled by μ. The class distance function, $g_a(\mathbf{x_n})$, is defined in terms of the model-specific document similarity measure, $SIM(d_1, d_2)$ - the exponential is arbitrarily used here to convert a similarity into a distance. Then, GPD theory states that the empirical expected loss function can then be minimized by using the iterative gradient-descent update equation

$$\mathbf{\Lambda}(t + 1) = \mathbf{\Lambda}(t) - \epsilon(t)\nabla l_k(\mathbf{x_n}) \tag{8}$$

where the modelset is updated once for each training example in $\mathbf{x_n} \in \mathbf{X}$ at each iteration t, and $\epsilon(t)$ is a monotonically decreasing function. If model parameters are considered independent, then using equations 5-8, it can be shown that document model, $\lambda^{\mathbf{b}}$, can be minimized using the per-model-parameter update equation:

$$\lambda^b_j(t + 1) \quad = \quad \lambda^b_j(t) - \epsilon(t)\frac{\partial l_a(\mathbf{x_n})}{\partial \lambda^b_j} \tag{9}$$

where

$$\frac{\partial l_a(\mathbf{x_n})}{\partial \lambda^b_j} \quad = \quad \alpha l_a(\mathbf{x_n}) \left\{1 - l_a(\mathbf{x_n})\right\} \frac{\partial d_a(\mathbf{x_n})}{\partial \lambda^b_j} \tag{10}$$

$$\frac{\partial d_a(\mathbf{x_n})}{\partial \lambda^b_j} \quad = \quad -\gamma g_b(\mathbf{x_n}) \times \frac{\partial SIM(\lambda^{\mathbf{a}}, \mathbf{x_n})}{\partial \lambda^j_b} \times \Psi_{ab} \tag{11}$$

$$\Psi_{ab} \quad = \quad \begin{cases} 1, & a = b \\ -V_{ab}, & \text{else} \end{cases} \tag{12}$$

$$V_{ab} \quad = \quad \frac{1}{C'}\left[\frac{1}{C'}\sum_{c=1}^{A}\left[\frac{g_b(x)}{g_c(x)}\delta(\mathbf{x_n} \notin C_c)\right]^{\mu}\right]^{-(1+\mu)/\mu} \tag{13}$$

$$C' \quad = \quad \sum_{c=1}^{A} \delta(\mathbf{x_n} \notin C_c) \tag{14}$$

Model training can then be performed using the following procedure. First a training set is constructed using a set of target document word vectors, $\mathbf{R} = (\mathbf{r_1}, \ldots, \mathbf{r_D})$, a set of training document word vectors, $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_N}\}$, and the $D \times X$ document similarity matrix, Ψ, where Ψ_{ij} is the binary function indicating whether target document i is similar to training document j. An initial set of target document models, $\mathbf{\Lambda}$, is bootstrapped appropriately. Then, for each training iteration, t, each training document, n is processed as follows:

1. The set of related documents, $\phi = \{\phi_1, \ldots, \phi_S\}$ is computed by selecting all documents where $\Psi_{sn} = 1$.
2. For each related document, ϕ_s in ϕ
   (a) Assume $\mathbf{x_n}$ belongs to class C_{ϕ_s} ie. $\mathbf{x_n}$ is an example of document ϕ_s.
   (b) Then, for each model $\lambda^{\mathbf{i}}$ in $\mathbf{\Lambda}$, the new model parameters, $\lambda'^{i}_{j}(t,s)$, are computed using equation 9.
3. Individual document model parameters are then updated to be the average of all individual model parameter updates, using $\lambda^i_j(t+1) = \frac{1}{S}\sum_{s=1}^{S} \lambda'^{i}_{j}(t,s)$

Iterative training continues until the change in the total empirical loss function, $L(\mathbf{\Lambda})$, is negligible.

3.1. Applying the PLSA model structure

The derivation above demonstrates how the ensemble of document models, $\mathbf{\Lambda}$, can be trained discriminatively, given a modeling architecture, a document similarity measure, $SIM(d_1, d_2)$, and a document relationship function that allows computation of $\delta(\mathbf{x_n} \notin C_c)$. For PLSA, SIM_{PVSM} in equation 3 can be used for $SIM(d_1, d_2)$, while the document relationship function can be derived from training document similarity annotations.

Defining the modeling architecture however requires a modeling assumption. Here, it is proposed that an individual document can be modeled using $\lambda^{\mathbf{i}}$ with $\lambda^i_j = d^i_j \gamma^i_j$, where d^i_j is the EFF term. γ^i_j is an importance weight that reflects the importance of the factor j for discrimination of document i. These factor-importance weights are similar to global term weights, such as IDF weights, but here, the importance weight is trained on a per-document basis. Conceptually, training such a model is equivalent to simply adjusting the position of the EFF document vectors in factor space to maximize discriminability.

The training procedure in section 3 can then be used with the following modifications to refine the PLSA model: 1) the initial model set is bootstrapped using standard ML PLSA to estimate $P(z_j|d_i)$, and setting $\gamma^i_j = 1$ for all models, 2) each training document $\mathbf{x_n}$ is folded into the factor space using one of the folding approaches described in section 2.1 and 3) the SIM_{PVSM} similarity measure is modified to include the γ weights, resulting in the weighted VSM similarity:

$$SIM_{GVSM}(\mathbf{d^1}, \mathbf{d^2}) = \frac{\gamma^{\mathbf{1}} \cdot \mathbf{d^1} \cdot \mathbf{d^2}}{|\gamma^{\mathbf{1}} \cdot \mathbf{d^1}|\,|\mathbf{d^2}|} \qquad (15)$$

4. EXPERIMENTS AND RESULTS

The proposed discriminative training process was evaluated on the TDT2 speech corpus. Experiments measured the performance of a discriminatively trained PLSA modelset, as well as the performances of a number of standard baseline systems.

Method	RDR mAP	QDR mAP
TFIDF	53.9	57.6
PVSM	37.0	29.9
TFIDF + PVSM	52.0	63.5
GVSM	66.0	41.4
TFIDF + GVSM	72.0	66.3

Table 1. Mean Average Precision results for the RDR and QDR tasks evaluated on the TDT2 corpus

4.1. Experiment setup

A 68 hour subset of speech was selected as the target document set for all experiments. TDT2 automatic speech recognition transcripts were used to create document vectors for each of these documents. Training and evaluation sets were then constructed for two tasks. The primary task, called Related Document Retrieval (RDR), was designed to evaluate shared topic open document set retrieval tasks, such as content-based recommendation and related document browsing. The document similarity matrix for GPD training was built using topic annotations from 91 standard topic groups in the TDT2 corpus. Documents were marked as related if they shared at least one topic label. The RDR evaluation set was constructed by selecting 41 additional query documents from TDT2 that shared a topic label with at least one document in the target document set. On average, each query had 15 related documents in the target database.

The second task, called Query-by-Document Retrieval (QDR), was designed to evaluate open topic open document set retrieval tasks, such as document-as-a-query information retrieval. The same process used for RDR was used to construct the QDR training and evaluation sets, except that topic exclusivity was enforced between the training and evaluation sets. Thus the QDR similarity matrix was constructed using only 36 topic groupings from the TDT2 training split while the evaluation set was constructed from 21 query documents (average of 11.5 related documents per query) that only had topics from the development and evaluation splits of TDT2.

Performance was evaluated for 3 baseline systems, 1) TFIDF: VSM using TFIDF with log TF tapering, 2) PVSM: ML-trained PLSA model with SIM_{PVSM} similarity metric, and 3) the fused system TFIDF+PVSM: linear fusion of the TFIDF and PVSM systems. All PLSA systems used 200 factors and probabilistic EFF predication for query folding. Fusion was performed using simple linear fusion: $SIM_{fused} = \rho SIM_{VSM} + (1-\rho) SIM_{PVSM}$, where ρ was coarsely tuned to maximize performance on a development set.

A maximum of 20 iterations was performed during discriminative training. This was sufficient to achieve a stable loss function, although generally, close to optimal performance was achieved in 5-10 iterations.

4.2. Related Document Recommendation

Evaluations were performed for the RDR task to compare the performances of a GPD-trained PLSA (GVSM) system and a fused GPD-trained system (TFIDF+GVSM) with the 3 baseline systems. The Mean Average Precision (mAP) results for these experiments are shown in table 1 and a precision-recall DET plot is shown in figure 1. The results demonstrated that the GPD-trained systems were clearly superior to the baseline systems, with absolute mAP gains of 12.1% and 18.1% for GVSM and TFIDF+GVSM respectively over the best baseline TFIDF system. Particularly pleasing was the dramatic 29% absolute mAP gain of GVSM over the standard PLSA

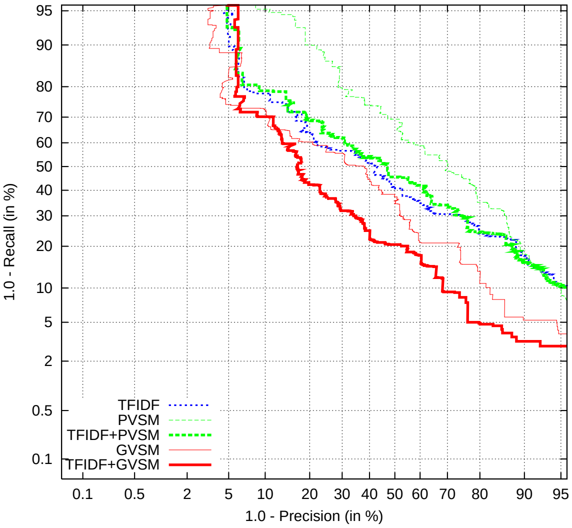

Fig. 1. Precision-Recall DET plot for RDR task

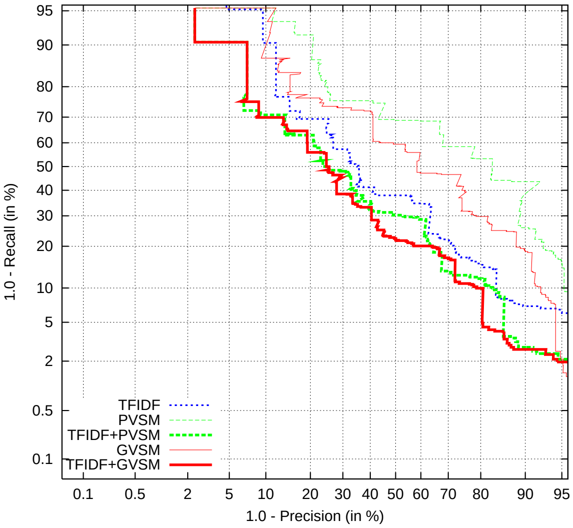

Fig. 2. Precision-Recall DET plot for QDR task

PVSM baseline system. This clearly demonstrates that the discriminative training approach was able to considerably improve document similarity performance for RDR while still preserving the ability to generalize to unseen documents.

4.3. Query-by-Document Spoken Document Retrieval

Experiments for the QDR task also demonstrated improvements for the GPD-trained systems, although the gains were more moderate. The mAP results are shown in table 1 and a precision-recall DET is shown in figure 2. Here, the fused TFIDF+GVSM was the best system, achieving an mAP of 66.3%. However, GVSM did not outperform TFIDF. Nevertheless, GVSM still achieved a significant absolute mAP gain of 11.5% over the baseline PLSA PVSM system.

The lower gains for QDR are a result of two factors. Firstly, QDR was clearly a much harder task *for PLSA*, demonstrated by the fact that performance of all PLSA systems was lower for QDR than RDR, even though TFIDF performance was better for QDR compared to RDR. Both PVSM and GVSM performance could potentially be improved by increasing the amount of training data or the number of semantic factors. Secondly, the GVSM training document similarity matrix for QDR was considerably more sparse than that used for RDR, since in this case, the matrix was only constructed from annotations for 36 topics (as opposed to 91 topics for RDR). Thus the amount of training data for GVSM was considerably lower for the QDR task, which may explain why the gains over PVSM were less. Nevertheless, it is clear that the discriminative training process is able to train a better set of models for generalizing to both unseen topics and unseen documents. This improvement in generalization is clearly demonstrated by the 2.8%/8.7% mAP gains observed for TFIDF+GVSM system over the TFIDF+PVSM and TFIDF systems respectively.

5. CONCLUSION

This work has presented a novel approach to improving spoken document similarity by using discriminatively trained document similarity models. Related document and query-by-document retrieval evaluations on the TDT2 corpus demonstrated significant gains in mAP over TFIDF and PLSA baseline systems. The experiments highlight that the proposed method improves the ability of document similarity models to generalize to both unseen documents and unseen topics, making it suitable for a variety of spoken document retrieval and similarity tasks.

6. REFERENCES

[1] C. D. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*, chapter Topics in Information Retrieval, pp. 539–41, MIT Press, 1999.

[2] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman, “Indexing by latent semantic analysis,” *Journal of the American Society of Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[3] Thomas Hofmann, “Probabilistic latent semantic analysis,” in *Proc. of Uncertainty in Artificial Intelligence, UAI’99*, Stockholm, 1999.

[4] A. Biem, “Minimum classification error training for online handwriting recognition,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2006.

[5] J. A. Lasserre, C. M. Bishop, and T. P. Minka, “Principled hybrids of generative and discriminative models,” in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 2006.

[6] Biing-Hwang Juang, Wu Hou, and Chin-Hui Lee, “Minimum classification error rate methods for speech recognition,” in *Speech and Audio Processing, IEEE Transactions on*, 1997.

[7] D. Povey and P.C. Woodland, “Minimum phone error and i-smoothing for improved discriminative training,” in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. IEEE International Conference on*, 2002.

[8] S. Katagiri, Biing-Hwang Juang, and Chin-Hui Lee, “Pattern recognition using a family of design algorithms based upon the generalized probabilistic descent method,” *Proceedings of the IEEE*, 1998.