A UNIFIED VIEW FOR DISCRIMINATIVE OBJECTIVE FUNCTIONS BASED ON NEGATIVE EXPONENTIAL OF DIFFERENCE MEASURE BETWEEN STRINGS

Atsushi Nakamura † *Erik McDermott* † *Shinji Watanabe* † *and Shigeru Katagiri* ‡

† NTT Communication Science Laboratories, NTT Corporation, Japan
‡ Faculty of Science and Engineering, Doshisha University, Japan
{ats, mcd, watanabe}@cslab.kecl.ntt.co.jp, skatagir@mail.doshisha.ac.jp

ABSTRACT

This paper presents a novel unified view of a wide variety of objective functions suitable for discriminative training applied to sequential pattern recognition problems, such as automatic speech recognition. Focusing on a central component of conventional objective functions, the sum of modified joint probabilities of observations and strings, the analysis generalizes these objective functions by weighting each term in the sum by an important function, the negative exponential of difference measure between strings. The interesting and valuable results of this investigation are highlighted in a comprehensive relationship chart that covers all of the common approaches (Maximum Mutual Information, Minimum Classification Error, Minimum Phone/Word Error), as well as corresponding novel generalizations and modifications of those approaches.

Index Terms— Automatic speech recognition, discriminative training, generalized objective function, negative exponential function, Laplace-Stieltjes transform

1. INTRODUCTION

It is currently well known that discriminative training can effectively improve the performance of sequential pattern recognition [1], as exemplified by state of the art speech recognition. Various approaches to discriminative training have been studied in order to increase performance. Common successful approaches include Maximum Mutual Information (MMI) [2], Minimum Classification Error (MCE) [3], Minimum Phone/Word Error (MPE/MWE) [4], and their modifications.

Each approach is primarily characterized by its definition of objective function. For example, the MMI objective function is defined so as to maximize the mutual information between observations and their corresponding strings, the MCE objective function directly addresses the notion of discriminating between correct and incorrect strings, and the MPE/MWE objective function is intended to minimize classification error weighted by an accuracy/error count that is measured at an arbitrary grain size, such as word, phoneme, phoneme-frame pair [5], etc.

Many of these objective functions become very similar to each other in the context of large vocabulary continuous speech recognition (LVCSR) [6]. For instance, comparing MMI and MCE with a linear loss function, the major implementational difference only concerns whether or not a competitor lattice or N-best list can include the correct string. Here we re-examine conventional formulations of discriminative objective functions. The result of this study is not only a new perspective on the commonality among the formulations, but the derivation of original and, we believe, promising approaches to discriminative training.

The novel unified view described in this study is based on a family of element functions, in terms of which most conventional objective functions and their modifications are commonly constructed. This analysis establishes the existence of a broader class of objective functions, allowing the construction of novel criteria using the same element functions with slightly different settings. An interesting result from these investigations is that the Laplace-Stieltjes transform (LST) of the cumulative error-count distribution can be naturally derived from a fractional form of the element functions; the LST can then be used to generate a generalized function family that includes the MPE/MWE objective function. The essential results of this comprehensive new framework are highlighted in a relationship chart that covers MMI, MCE, MPE/MWE, as well as corresponding novel generalizations and modifications of those approaches (see Figure 2).

This paper is organized as follows: Section 2 introduces a family of element functions based on the negative exponential function of difference measure. Section 3 proves that simple fractional forms of element functions can construct most conventional objective functions. Section 4 demonstrates that novel generalized objective functions can also be constructed in terms of the element functions.

2. NEGATIVE EXPONENTIAL OF DIFFERENCE MEASURE AND AN ELEMENT FUNCTION FAMILY

We consider a set of strings of symbols, such as linguistic symbols (phonemes, words, etc.), and introduce a negative exponential function of the difference measure $\Delta(S_1, S_2)$ between two strings S_1 and S_2 with the exponential decay factor σ:

$$\exp(-\sigma\Delta(S_1, S_2)). \quad (1)$$

Most typically, the symbol edit distance (a.k.a. Levenshtein distance) between strings is applicable to $\Delta(S_1, S_2)$. When S_1 and S_2 are the "correct" and "incorrect" strings, respectively, the edit distance can be regarded as the "error count." Assuming that the two strings are associated with a common sequence of observation units, a type of symbol-unit error count, e.g. "phone frame error" count [5], can also be used. Although this paper focuses on error count related measures, the difference measure, in general, does not necessarily have to correspond directly to the

error count. We can interpret Eq. (1) as meaning that the opposite of difference measure, which thus corresponds to similarity, is mapped into a probability-like (0,1] measure (when $\sigma > 0$).

Now, we take the case of LVCSR, and assume there is a pair (X_r, S_r) of an acoustic observation sequence and its correctly transcribed string, and a lattice or N-best represented set $\mathcal{S} = \{S_k \mid k = 1, 2, 3, \cdots\}$ of speech-recognized strings for X_r . We can then define a family of σ-parameterized functions by the sum of modified joint probability densities of X_r and S_k , each term of which is weighted by the negative exponential of difference measure between correct and recognized strings,

$$\begin{aligned} \psi_\sigma(X_r, S_r) &= \sum_k P_{\Lambda L}(S_k)^{\eta\phi}\, p_{\Lambda A}(X_r \mid S_k)^{\phi} \exp(-\sigma\Delta(S_r, S_k)) \\ &= \sum_k p_\Lambda(X_r, S_k) \exp(-\sigma\Delta_k), \end{aligned} \tag{2}$$

where Λ , η and ϕ are a set of acoustic (Λ_A) and language (Λ_L) model parameters, scaling factor for $P_{\Lambda L}(S_k)$, and density smoothing factor, respectively, and $\Delta(S_r, S_k)$ is shortened into Δ_k .

This simple family of functions, in fact, forms a principal set of elements for several types of discriminative objective functions. Henceforth, we refer generically to each member function of this family as ψ-probability or psi-probability, standing for "pseudo (=$\psi\varepsilon\upsilon\delta o$) probability."

3. REPRESENTATION OF CONVENTIONAL OBJECTIVE FUNCTIONS USING PSI-PROBABILITIES

In this section, we prove that several types of conventional discriminative objective functions, including MMI, MCE, MPE/MWE, and well-known modifications thereof, can be constructed using ψ-probabilities. In the following, all types of objective functions represented with ψ-probabilities are defined to be maximized, for the sake of formal consistency. Namely, some functions in the formulations might be negated or inverted if they are originally defined to be minimized. Also note that the objective functions in this paper are defined w.r.t. a single observation sequence, and that each criterion can be immediately extended to deal with an overall set of observation sequences.

3.1. Maximum Mutual Information (MMI)

MMI training [2] maximizes the mutual information between X_r and S_r by means of the objective function,

$$\log \frac{p_\Lambda(X_r, S_r)}{\sum_k p_\Lambda(X_r, S_k)} = \log F_{MMI}(X_r, \Lambda). \tag{3}$$

Note that, as defined in Eq. (2), the scaling and smoothing factors are already involved in $p_\Lambda(X_r, S_k)$. We can then describe the function equivalent to $F_{MMI}(X_r, \Lambda)$ using two extreme expressions of ψ-probability:

$$\psi_0 = \sum_k p_\Lambda(X_r, S_k) \exp(-0 \cdot \Delta_k) = \sum_k p_\Lambda(X_r, S_k) \tag{4}$$

and

$$\psi_\infty = \lim_{\sigma\to\infty} \sum_k p_\Lambda(X_r, S_k) \exp(-\sigma\Delta_k) = p_\Lambda(X_r, S_r) \tag{5}$$

as

$$\frac{\psi_\infty}{\psi_0} = \frac{p_\Lambda(X_r, S_r)}{\sum_k p_\Lambda(X_r, S_k)} \equiv F_{MMI}(X_r, \Lambda). \tag{6}$$

This formulation can easily be extended to the "boosted" version of MMI (Boosted MMI [7]):

$$\frac{\psi_\infty}{\psi_{(-\sigma)}} = \frac{p_\Lambda(X_r, S_r)}{\sum_k p_\Lambda(X_r, S_k) \exp(\sigma\Delta_k)} \equiv F_{bMMI}(X_r, \Lambda). \tag{7}$$

3.2. Minimum Classification Error (MCE)

MCE training is directly aimed at discriminating S_r from the set of incorrect strings, all of which are elements of the overall set $\mathcal{S}$ of recognized strings. The framework of MCE involves a variety of formulations for objective functions. We here assume one of the formulations of misclassification measure $d_r(X_r, \Lambda)$ suitable for LVCSR [3], defined in terms of log-likelihood based discriminant functions $\mathcal{G}_k(X_r, \Lambda) = \log(P_{\Lambda L}(S_k)^{\eta}\, p_{\Lambda A}(X_r \mid S_k))$, and define $F_{MCE}(X_r, \Lambda)$ as a core part in the formulation of the misclassification measure,

$$\begin{aligned} d_r(X_r, \Lambda) &= -\mathcal{G}_r(X_r, \Lambda) + \log\left(\frac{1}{C} \sum_{k\mid\Delta_k\neq 0} e^{\mathcal{G}_k(X_r, \Lambda)\cdot\phi} \right)^{\frac{1}{\phi}} \\ &= \frac{1}{\phi}\left(\log \frac{\sum_{k\mid\Delta_k\neq 0} p_\Lambda(X_r, S_k)}{p_\Lambda(X_r, S_r)} - \log C \right) \\ &= \frac{1}{\phi}\left(\log F_{MCE}(X_r, \Lambda) - \log C \right), \end{aligned} \tag{8}$$

where C is the cardinal number of a set $\mathcal{S} - \{S_r\}$. The final form of the MCE objective function is realized by a (possibly linear) loss function that takes $d_r(X_r, \Lambda)$ as its input. The direct expression of the MCE objective function cannot be derived by using the ψ-probabilities alone. Instead, the function equivalent to the inverse of $F_{MCE}(X_r, \Lambda)$, a core part in $d_r(X_r, \Lambda)$, is derived as

$$\frac{\psi_0}{\psi_\infty - \psi_0} = \frac{p_\Lambda(X_r, S_r)}{\sum_k p_\Lambda(X_r, S_k) - p_\Lambda(X_r, S_r)} \equiv \frac{1}{F_{MCE}(X_r, \Lambda)}. \tag{9}$$

3.3. Minimum Phone/Word Error (MPE/MWE)

MPE/MWE training [4] minimizes classification error weighted by an accuracy/error count that is measured at an arbitrary grain size, such as word, phoneme, phoneme-frame pair, etc. Although the original definition of the MPE/MWE objective function is based on a raw accuracy count, we here employ another equivalent formulation, based on a difference measure, such as the raw error count,

$$F_{MPE}(X_r, \Lambda) = \frac{\sum_k p_\Lambda(X_r, S_k)\Delta_k}{\sum_k p_\Lambda(X_r, S_k)}, \tag{10}$$

or, in other words, the model-based expectation of error count over the set $\mathcal{S}$ of recognized strings. This expectation can also be seen as the pivot for MPE/MWE learning based on the derivative of the expression: S_k will be treated positively/negatively if Δ_k is smaller/larger than the expected error [8]. Even though Eq. (10) explicitly involves the linear count of errors, which is entirely

absent from the MMI and MCE formulations, the ψ-probability based formulation can cover such a case by introducing the partial derivative of ψ-probability w.r.t. σ ,

$$\psi'_{\sigma} = \left.\frac{\partial \psi_{\upsilon}}{\partial \upsilon}\right|_{\upsilon=\sigma} = -\sum_{k} p_{\Lambda}(X_r, S_k) e^{-\sigma\Delta_k} \Delta_k \qquad (11)$$

as

$$\frac{\psi'_0}{\psi_0} = -\frac{\sum_k p_{\Lambda}(X_r, S_k)\Delta_k}{\sum_k p_{\Lambda}(X_r, S_k)} \equiv -F_{MPE}(X_r, \mathbf{\Lambda}), \qquad (12)$$

and its boosted version (Boosted MPE)

$$\frac{\psi'_{(-\sigma)}}{\psi_{(-\sigma)}} = -\frac{\sum_k p_{\Lambda}(X_r, S_k)\exp(\sigma\Delta_k)\Delta_k}{\sum_k p_{\Lambda}(X_r, S_k)\exp(\sigma\Delta_k)} \equiv -F_{bMPE}(X_r, \mathbf{\Lambda}). \qquad (13)$$

4. GENERALIZED VERSIONS OF DISCIMINATIVE OBJECTIVE FUNCTIONS

As proved in the previous section, ψ-probabilities can be used to formulate several common discriminative objective functions. This section describes natural extensions of the ψ-probability based formulations, leading to novel generalized objective functions. The conventional and generalized functions discussed in this paper are summarized in a comprehensive relationship chart.

4.1. Generalized MMI/MCE and Maximum String Similarity

The negative exponential weight in ψ-probability decays $p_{\Lambda}(X_r, S_k)$ according to the difference measure with decay rate σ . The arbitrary choice of σ produces several differently decaying ψ-probabilities. We can derive some generalized objective functions by making use of these characteristics.

For example, a fractional form, a rapidly decaying ψ-probability over a slowly decaying or growing one, provides a generalized type of MMI objective function:

$$\frac{\psi_{\sigma_1}}{\psi_{\sigma_2}} = \frac{\sum_k p_{\Lambda}(X_r, S_k) e^{-\sigma_1\Delta_k}}{\sum_k p_{\Lambda}(X_r, S_k) e^{-\sigma_2\Delta_k}} \quad (\sigma_1 > \sigma_2). \qquad (14)$$

Here, the "correct" strings are defined in a loose, flexible sense, as the weighted sum in the numerator also includes the densities for "nearly correct" strings. And, in the denominator, the densities are boosted/reduced for strings that are not even close. In the limit of $\sigma_1 \to \infty$, Eq. (14) approaches the ordinary ($\sigma_2 = 0$) or Boosted ($\sigma_2 = -\sigma < 0$) MMI (Eqs. (6) and (7)). A similar generalization can be also made for MCE:

$$\frac{\psi_{\sigma_1}}{\psi_{\sigma_2} - \psi_{\sigma_3}} = \frac{\sum_k p_{\Lambda}(X_r, S_k) e^{-\sigma_1\Delta_k}}{\sum_k p_{\Lambda}(X_r, S_k) e^{-\sigma_2\Delta_k} - \sum_k p_{\Lambda}(X_r, S_k) e^{-\sigma_3\Delta_k}} \quad (\sigma_1 \simeq \sigma_3 > \sigma_2). \qquad (15)$$

In the denominator of Eq. (15), the densities for the loosely defined "correct" strings are subtracted from the boosted/reduced sum. The expression approaches MCE (Eq. (9)) in the limit of $\sigma_1 \to \infty$ and $\sigma_3 \to \infty$ with $\sigma_2 = 0$. Previous work by the authors has demonstrated the benefits of using a *set* of correct strings rather than a single correct string [3]. The above generalizations can reasonably be expected to yield similar benefits. The effect of exponential weight in these generalized objective functions is depicted in Figure 1.

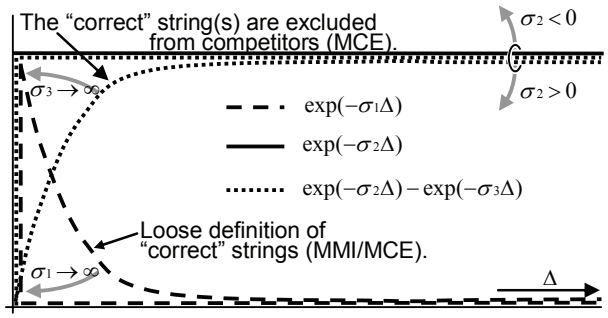

Figure 1. The effect of negative exponential weight in generalized MMI and MCE objective functions.

Another interpretation of Eq. (14) is provided by rewriting it as

$$\frac{\sum_k p_{\Lambda}(X_r, S_k) e^{-\sigma_1\Delta_k}}{\sum_k p_{\Lambda}(X_r, S_k) e^{-\sigma_2\Delta_k}} = \frac{\sum_k p_{\Lambda}(X_r, S_k) e^{\sigma'\Delta_k} e^{-\sigma\Delta_k}}{\sum_k p_{\Lambda}(X_r, S_k) e^{\sigma'\Delta_k}} \quad (\sigma = \sigma_1 - \sigma_2,\ \sigma' = -\sigma_2). \qquad (16)$$

Setting $\sigma' = 0$ immediately leads to

$$\frac{\psi_{\sigma}}{\psi_0} = \frac{\sum_k p_{\Lambda}(X_r, S_k)\exp(-\sigma\Delta_k)}{\sum_k p_{\Lambda}(X_r, S_k)}. \qquad (17)$$

This corresponds to the expectation of the negative exponential of difference measure, the maximization of which maximizes the expectation of a similarity measure between correct and recognized strings (Maximum String Similarity), using ordinary (Eq. (17)) and boosted (Eq. (16)) densities. Since $1 - \exp(-\sigma\Delta_k)$ becomes close to $\sigma\Delta_k$ for a sufficiently small σ , Maximum String Similarity is expected to work similarly to minimizing the expectation of difference measure, i.e. MPE/MWE. In general, the concavity of $-\exp(-\sigma\Delta_k)$ lowers the expected error (the pivot for positive/negative string handling during learning) compared to conventional MPE/MWE (see Section 3.3)

As regards the computation of these functions over a lattice, the ψ-probability for each arc/node can be computed via the ordinary forward-backward algorithm, as long as the negative exponential of difference measure for each string in the lattice can be factorized into a local factor for each arc/node. Numerical lattice subtraction [6] can be used to implement the Generalized MCE criterion proposed in Eq. (15).

We also note that, in the exponential function based formulations presented in this section, we may adaptively set the decay factor σ , which closely relates to novel interpretations of the large margin approach that have recently been studied [9,10].

4.2. Generalized Minimum Error Moment

Eq. (17) also equals the Laplace-Stieltjes transform (LST) of a cumulative distribution of the difference measure, e.g. error count, over the set $\mathcal{S}$ of recognized strings:

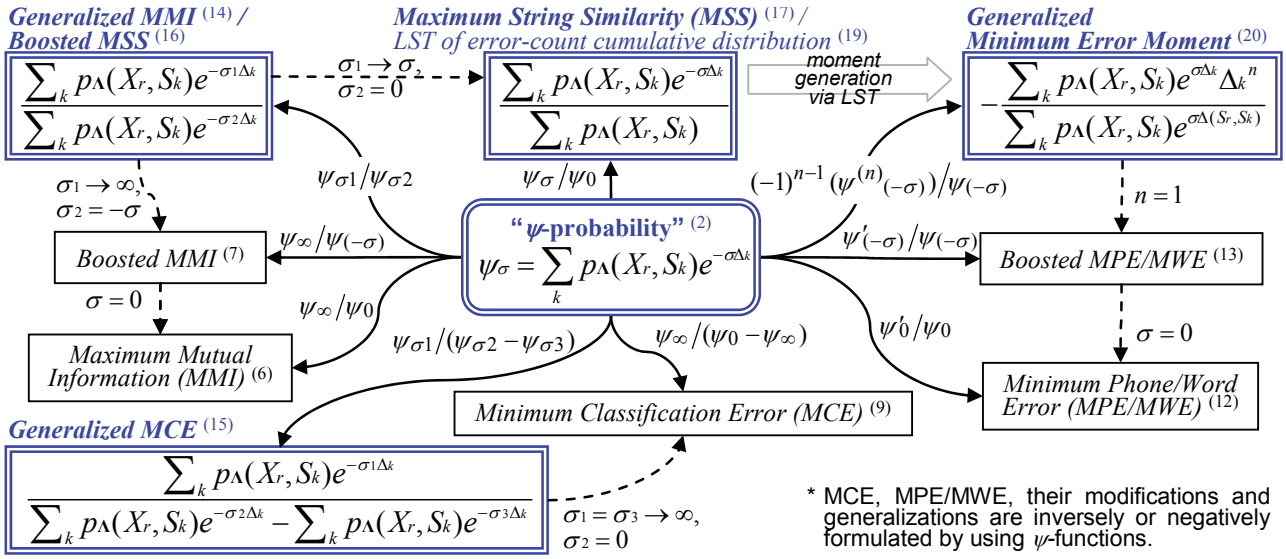

Figure 2. A comprehensive relationship among discriminative objective functions. Double-lined boxes denote novel functions proposed in this paper. Each number in parenthesis provides a link to the corresponding equation in the body text.

$$\chi(\delta) = \frac{\sum_{k|\Delta_k \le \delta} p_\lambda(X_r, S_k)}{\sum_k p_\lambda(X_r, S_k)}, \tag{18}$$

$$\begin{aligned} \mathrm{LST}[\chi(\delta)] &= \int_0^\infty e^{-\sigma\delta} d\chi(\delta) \\ &= \frac{\sum_k p_\lambda(X_r, S_k)\exp(-\sigma\Delta_k)}{\sum_k p_\lambda(X_r, S_k)} \equiv \frac{\psi_\sigma}{\psi_0}. \end{aligned} \tag{19}$$

By means of the property of LST, ψ_σ/ψ_0 can generate the origin moment of an arbitrary order n, leading to a generalized objective function, the maximization of which minimizes the (boosted) moment of an arbitrary order n (Generalized Minimum Error Moment):

$$(-1)^{n-1}\frac{\psi^{(n)}(-\sigma)}{\psi(-\sigma)} = -\frac{\sum_k p_\lambda(X_r, S_k) e^{\sigma\Delta_k}\Delta_k{}^n}{\sum_k p_\lambda(X_r, S_k) e^{\sigma\Delta_k}}. \tag{20}$$

In the case of $n=1$, Eq. (20) equals the ordinary ($\sigma = 0$) or boosted ($\sigma > 0$) MPE/MWE (Eqs. (12) and (13)). Setting $n > 1$ enables a novel type of discriminative training that uses a higher order moment over a lattice or an N-best list.

As for the computation of Eq. (20) over a lattice, foundational work with MPE/MWE [4] has presented an elegant solution to this for the specific case of $n=1$, where $\Delta_k{}^n = \left(\Delta(S_r, S_k)\right)^n$ for each S_k in the lattice can be additively partitioned into local components for each arc/node. On the other hand, recent work by the authors proposing an error-indexed forward-backward algorithm [8] enables the computation for an arbitrary order n.

Figure 2 illustrates the relationships between the conventional and generalized discriminative functions, summarizing the novel unified view centered on the use of ψ-probability, itself based on the negative exponential of string difference.

5. CONCLUSION

A novel unified view for discriminative objective functions has been presented here. A family of element functions based on the negative exponential of difference measure has been proved to broadly cover conventional objective functions, as well as their generalizations. Future work will include a series of experimental evaluations on the generalized functions proposed in this paper.

6. ACKNOWLEDGMENT

The research described in this paper is partially supported by the Grant-in-Aid for Scientific Research No. 19300064, Japan Society for the Promotion of Science.

7. REFERENCES

[1] X. He, et al., "Discriminative learning in sequential pattern recognition," *IEEE SP Mag.* 25, 5, pp. 14-36, September 2008.

[2] D. Povey, Discriminative training for large vocabulary Speech Recognition, *Ph.D. thesis, Cambridge University,* 2004.

[3] E. McDermott, et al., "Discriminative training for large vocabulary speech recognition using Minimum Classification Error," *IEEE Trans. ASLP*, 15, 1, pp. 203-223, January 2007.

[4]D. Povey and P. Woodland, "Minimum Phone Error and I-smoothing for improved discriminative training," in *Proc. IEEE ICASSP*, pp. 105-108, 2002.

[5] J. Zheng and A. Stolcke, "Improved discriminative training using phone lattices," in *Proc. Interspeech*, pp. 2125-2128, 2005.

[6] W. Macherey, et al., "Investigations on error minimizing training criteria for discriminative training in automatic speech recognition," in *Proc. Eurospeech*, pp. 2133-2136, 2005.

[7] D. Povey, et al., "Boosted MMI for model and feature-space discriminative training," in *Proc. IEEE ICASSP*, pp. 4057-4060, 2008.

[8] E. McDermott and A. Nakamura, "Flexible discriminative training based on equal error group scores obtained from an error-indexed forward-backward algorithm," in *Proc. Interspeech*, pp. 2398-2401, 2008.

[9] G. Heigold, et al., "Modified MMI/MPE: A direct evaluation of the margin in speech recognition," In *Proc. ICML*, pp. 384-391, 2008.

[10] G. Saon and D. Povey, "Penalty function maximization for large margin HMM training," In *Proc. Interspeech*, pp. 920-923, 2008.