# A NOVEL DECISION FUNCTION AND THE ASSOCIATED DECISION-FEEDBACK LEARNING FOR SPEECH TRANSLATION

*Yaodong Zhang*

MIT CSAIL
Cambridge, Massachusetts 02139, USA
ydzhang@csail.mit.edu

*Li Deng, Xiaodong He, Alex Acero*

Microsoft Research
Redmond, Washington 98052, USA
{deng,xiaohe,alex.acero}@microsoft.com

**ABSTRACT**

In this paper we report our recent development of an end-to-end integrative design methodology for speech translation. Specifically, a novel decision function is proposed based on the Bayesian analysis, and the associated discriminative learning technique is presented based on the decision-feedback principle. The decision function in our end-to-end design methodology integrates acoustic scores, language model scores and translation scores to refine the translation hypotheses and to determine the best translation candidate. This Bayesian-guided decision function is then embedded into the training process that jointly learns the parameters in speech recognition and machine translation sub-systems in the overall speech translation system. The resulting decision-feedback learning takes a functional form similar to the minimum classification error training. Experimental results obtained on the IWSLT DIALOG 2010 database showed that the proposed system outperformed the baseline system in terms of BLEU score by 2.3 points.

***Index Terms***— speech translation, decision feedback, integrative design, discriminative training

## 1. INTRODUCTION

Speech translation aims at automatically converting speech input in one language to text output in another language. A straightforward way of accomplishing speech translation is to build a two-stage system [10, 12], which combines the state-of-the-art techniques from automatic speech recognition (ASR) and statistical machine translation (SMT). To improve the translation performance, since there could be multiple recognition hypotheses available from n-best or lattice, the SMT engine can be designed to accept multiple input and make a decision based on some score fusion techniques [1, 6]. One problem of the direct combination or concatenative approach is the mismatch between the training corpus (clean) and the decoding environment (noisy) for the SMT system. In addition, errors made at the ASR stage cannot be recovered in the subsequent SMT stage.

From the point of view of the Bayesian optimal decision theory, the concatenative approach is clearly sub-optimal. To address this problem, in our earlier work on speech understanding, where the understanding module was a classifier trained on clean sentences with their corresponding semantic labels, we observed that the classification performance decreased vastly when the classifier's input was given with ASR output. A joint optimal design significantly improved the speech understanding performance [13]. Motivated by this earlier success, in this work, we formulate the speech translation problem as an optimal Bayesian decision problem. Specifically, we derive the optimal decision function which integrates acoustic scores, language model scores and translation scores together to determine the best translation candidate for a given speech input.

The decision function is then embedded into the training objective function, which enables joint learning of the parameters in the ASR and SMT sub-systems as the constituents of the overall speech translation system. The discriminative training strategy developed in this work is based on the decision-feedback principle, where the decision function that is used as the scoring function in the decoding process becomes a part of the optimization procedure of the entire system. As a result, the parameters in the speech translation system, including both ASR and SMT sub-systems, can be jointly learned by adjusting their current values so as to optimize the desired objective function or evaluation metrics. The optimization direction is guided by the feedback determined by the current set of parameters.

The rest of this paper is organized as follows. An overview of the proposed speech translation system design is provided in Section 2, which includes the derivation of the Bayesian decision function. In Section 3, the decision-feedback learning using the derived decision function is presented. The evaluation results are reported and analyzed in Section 4. Finally, we conclude this paper and provide plans for future work.

## 2. SPEECH TRANSLATION SYSTEM DESIGN

### 2.1. System Overview

Our speech translation system takes speech utterances as input, and outputs the translated utterances in text. Assuming the availability of existing ASR and SMT systems, the proposed speech translation system can be built in several steps. First, the existing ASR and SMT systems are concatenated to produce multiple translation hypotheses for each of the training speech utterances. Second, the translation hypotheses, together with translation references provided in the training data, are used to modify the parameters of the existing ASR and SMT systems using the algorithm described in Section 3. Third, the adjusted, new ASR and SMT systems are deployed to process each of the test speech utterances. Both ASR and SMT systems produce multiple hypotheses as the output, together with the associated acoustic scores, language model scores, and translation scores. The final best translated utterance is selected from the SMT hypotheses using the decision function that integrates these three sets of scores.

### 2.2. The Integrative Decision Function

The decision function is derived by using the Bayesian analysis. For concreteness and without loss of generality, we assume that our task is to translate from Chinese (C) speech utterances to English (E)

text. The most general form of Bayesian decision for this Chinese to English speech translation task can be represented as

$$\hat{E_r} = \arg\max_{E_r} P(E_r|X_r) \tag{1}$$

where $X_r$ is the speech input of $r^{th}$ Chinese utterance and $E_r$ is the corresponding English translation. By applying Bayesian analysis, the above equation can be expanded and approximated as

$$\begin{aligned}
\hat{E_r} &= \arg\max_{E_r} P(E_r|X_r) \\
&= \arg\max_{E_r} \sum_{C_r} P(E_r, C_r|X_r) \\
&= \arg\max_{E_r} \sum_{C_r} P(E_r|C_r, X_r)P(X_r|C_r)P(C_r) \\
&\approx \arg\max_{E_r} \sum_{C_r} P(E_r|C_r)P(X_r|C_r)P(C_r) \\
&\approx \arg\max_{E_r} \max_{C_r} P(E_r|C_r)P(X_r|C_r)P(C_r)
\end{aligned} \tag{2}$$

where $P(E_r|C_r)$ denotes the translation score given Chinese input $C_r$, $P(X_r|C_r)$ represents the acoustic score given $X_r$, and $P(C_r)$ is the language model score for the recognized Chinese utterance. Note that $P(E_r|C_r, X_r)$ is approximated by $P(E_r|C_r)$ by assuming that the translation is independent from the speech signal given the speech recognition hypothesis. Although the second approximation, replacing $\sum$ with $\max$ [10], may introduce inaccurate estimations, it greatly simplifies the development of the discriminative learning algorithm to be described in Section 3.

The product form of the three sets of the probabilities in Eq.2, or its equivalent form after taking the logarithm, constitutes the integrative decision function for the proposed speech translation system, which is what we will use to perform decision-feedback learning.

## 3. DECISION-FEEDBACK TRAINING

While the use of the integrative decision function for speech translation scoring takes into account the contributions from both ASR and SMT systems, the direct concatenative approach still has the deficiency that the parameters in the ASR and SMT systems are optimized towards their own respective instead of the combined, end-to-end speech to text translation performance. To overcome this deficiency, a decision-feedback learning technique is developed based on the minimum classification error (MCE) formulation [4, 2]. The MCE objective function is modified from the sentence recognition error rate, as developed originally for ASR, to the translation error rate as a measure of the speech translation quality.

### 3.1. The MCE Objective Function

To derive the objective function for optimization, we first define the class-discriminant function $D(\cdot)$ as

$$D(E_r, C_r; X_r) = \log\left[P(E_r|C_r)P(X_r|C_r)P(C_r)\right] \tag{3}$$

which acts as the scoring function for the classification decision mapping from the Chinese speech input $X_r$ to the corresponding recognition hypothesis $C_r$ and translation hypothesis $E_r$.

$$X_r \longrightarrow (E_r, C_r) \tag{4}$$

In the conventional MCE training, the correct classification hypothesis is the true reference. However, for SMT, the translation reference might not be achievable due to the limitations of the model [8]. Thus, similar to [8], the best possible translation hypothesis is selected as an approximation of the correct classification reference. Specifically, to judge the goodness of each classification decision $(E_r, C_r)$, we apply the commonly used BLEU evaluation metric on $E_r$ in each $(E_r, C_r)$ pair. Given the translation reference $R_r$, the BLEU scoring function is defined as BLEU$(R_r, E_r)$

Note that often times the BLEU score is calculated on the corpus level, while here we use a smoothed version of BLEU score defined on the sentence level [9]. Since the BLEU score averages the n-gram appearance in the test translation against reference translations, it is possible for different translation candidates $E_r^i$ to have the same BLEU score. As a result, the BLEU score is not enough to determine the best $(E_r, C_r)$ pair. To fix this, for decision pairs $(E_r^1, C_r^1)$ and $(E_r^2, C_r^2)$ when BLEU$(E_r^1)$ = BLEU$(E_r^2)$, the pair with higher $D(\cdot)$ score is chosen as the best decision. Formally, the correct classification decision pair $(E_r, C_r)$ for $X_r$ is determined in two steps. First, from the translation hypothesis set $T_r$, the best translated English sentence $E_r^0$ is selected by

$$E_r^0 = \arg\max_{E_r^i \in T_r} \text{BLEU}(R_r, E_r^i) \tag{5}$$

Second, from the recognition hypothesis set $S_r$, $C_r^0$ is selected by

$$C_r^0 = \arg\max_{C_r^i \in S_r} D(E_r^0, C_r^i; X_r) \tag{6}$$

Selecting the correct classification pair $(E_r, C_r)$ can be viewed as a two-key sorting process. Given all classification decision pairs $(E_r^i, C_r^i)$ for the input $X_r$, a ranked list can be build by considering the BLEU score of $E_r^i$ in each pair as the primary key and the $D(\cdot)$ score as the secondary key. The top ranked pair is the correct classification decision, while all remaining pairs are incorrect/competing decisions. To build the set of competing decision pairs $U_r$, the following rules are applied to ensure the correctness of the MCE training framework. Starting from going down the ranked list from the second decision pair (the first competing decision), every time a distinct translation hypothesis $E_r^i$ is seen, if its corresponding recognition $C_r^i$ is not included in any pairs already in $U_r$, this $(E_r^i, C_r^i)$ pair is added into $U_r$. If there is already a pair containing $C_r^i$ in $U_r$, this pair is ignored and the next pair is checked until all recognition hypotheses are consumed. Note that this procedure ensures that 1) $C_r^0$ is not equal to any of $C_r^i$ in competing pairs in $U_r$, which prevents the contribution cancellation in the MCE update if both the correct and competing decisions contain the same hypothesis; 2) every speech recognition hypothesis is used to provide discriminative information for the final translation decision, which is important because it is common that top translation hypotheses always come from the first one or two speech recognition hypotheses.

After collecting all competing pairs into the set $U_r$, the class-specific misclassification function $d_r(\cdot)$ can be defined to calculate the raw loss of the current classification decisions given $X_r$.

$$\begin{aligned}
d_r(X_r) = &- D(E_r^0, C_r^0; X_r) \\
&+ \log\left\{\frac{1}{|U_r|}\sum_{i}^{|U_r|}\exp\left[\eta D(E_r^i, C_r^i; X_r)\right]\right\}^{\frac{1}{\eta}}
\end{aligned} \tag{7}$$

This raw loss can be smoothed by the sigmoid function as

$$l_r(X_r) = \frac{1}{1+\exp(-\alpha d_r(X_r)+\beta)} \tag{8}$$

where $\eta, \alpha$ and $\beta$ are the standard parameters in the MCE training. Summing up loss from all training samples, the total loss $L$ is

$$L(A^{(t)}, LM_c^{(t)}, TM^{(t)}) = \sum_r l_r(X_r) \quad (9)$$

where $A^{(t)}, LM_c^{(t)}, TM^{(t)}$ is the current acoustic model, Chinese language model and translation model, respectively. Note that the translation model includes an English language model.

### 3.2. Parameter Estimation

#### *3.2.1. Language Model Update*

Both Chinese and English language models (LMs) are updated. Let $P(w_x|w_y)$ denote the bigram (for simplicity but without loss of generality) log-probability of the word $w_x$ given $w_y$. The English LM update can be derived by using the steepest descent method as

$$\begin{aligned} P(w_x|w_y)^{t+1} &= P(w_x|w_y)^t - \\ &- \epsilon\alpha \sum_r l_r(d_r(X_r))[1 - l_r(d_r(X_r))] \frac{\partial d_r(X_r)}{\partial P(w_x|w_y)} \end{aligned} \quad (10)$$

As described in Section 3.1, for each Chinese speech input $X_r$, the correct classification pair $(E_r^0, C_r^0)$ and the competing set $U_r$ can be obtained by sorting the recognition and translation pairs $(E_r^i, C_r^j)$. Let $N(E_r^i, w_x, w_y)$ denotes the number of times the bigram $w_x w_y$ appears in sentence $E_r^i$. The log-probability of the sentence $E_r^i$ can be then written as

$$\begin{aligned} P(E_r^i) &= \log \left[ \prod_{w_x, w_y} P(w_y|w_x)^{N(E_r^i, w_x, w_y)} \right] \\ &= \sum_{w_x, w_y} N(E_r^i, w_x, w_y) \log P(w_y|w_x) \end{aligned} \quad (11)$$

Given $(E_r^0, C_r^0)$ and the competing set $U_r$ for $X_r$, we can obtain the partial derivative of $d_r(X_r)$ for English bigrams as

$$\frac{\partial d_r(X_r)}{\partial P(w_x|w_y)} = -N(E_r^0, w_x, w_y) + \sum_i^{|U_r|} H_r^i N(E_r^i, w_x, w_y) \quad (12)$$

where

$$H_r^i = \frac{\exp[\eta D(E_r^i, C_r^i; X_r)]}{\sum_i^{|U_r|} \exp[\eta D(E_r^i, C_r^i; X_r)]} \quad (13)$$

is the weighting factor. Note that the Chinese LM update formula can be obtained by replacing every $E_r^i$ with its corresponding $C_r^i$.

#### *3.2.2. Phrase Table Update*

We consider here only the phrase-based translation models with non-parametric distributions of $P(E_r|C_r)$, which can be further decomposed into [5]:

$$P(E_r|C_r) = P^P(E_r|C_r) \cdot P^R(E_r|C_r) \cdot P(E_r) \cdot \omega^{L(E_r)} \quad (14)$$

where $P^P$ is the phrase translation probability, $P^R$ is the reordering probability and $\omega$ is the length penalty of the translation output. In this paper, we only focus on updating the phrase translation probabilities and leave the update of the other parts for future work. Given an alignment $A_r$ that aligns phrases in $E_r$ and $C_r$, the phrase translation probability of the sentence pair $E_r$ and $C_r$ is defined as [5]:

$$P^P(E_r|C_r) = \prod_{(p_e, p_c) \in A_r} P(p_e|p_c) P(p_c|p_e) L(p_e|p_c) L(p_c|p_e) \quad (15)$$

where $P(\cdot)$ is the phrase translation probability in both directions and $L(\cdot)$ is the lexicon weighting factor for phrases in both directions. According to the phrase extraction algorithm [5], the lexicon weighting factor is fixed when a particular phrase pair is extracted. As a result, we leave $L(\cdot)$ unchanged and only update the phrase translation probability $P(\cdot)$.

The phrase translation table can be viewed as a "bigram language model" capturing phrase co-occurrences between two languages. Thus, the updating formula (omitted due to the space limit) similar to the LM update can be derived for the phrase translation probabilities.

## 4. EVALUATION

### 4.1. Dataset and Evaluation Criterion

The experiments were carried out using the IWSLT DIALOG 2010 dataset which contains human-mediated dialogs in travel domain between Chinese and English. The dataset has two parts. The first part contains around 30,000 sentences in clean parallel text only, while the second part includes another 2,952 sentences with both ASR output and clean text. The ASR output is in the standard HTK lattice format. The dataset also provides the extracted 1-best and 20-best ASR hypotheses with word segmentation information.

In the current experiments, we only focus on translating from Chinese to English using Chinese ASR hypotheses as input. 296 and 294 out of 2,952 Chinese utterances were randomly selected as the development set and test set, respectively. The remaining 2,362 Chinese utterances were used for the decision-feedback training. Each Chinese utterance has 4 to 16 English translation references. The standard BLEU score was used as the evaluation criterion.

### 4.2. The Baseline System

The baseline system directly cascades the existing ASR and SMT systems without using the integrative decision function and the decision-feedback training. Since the dataset has already provided the 20-best ASR hypotheses for each Chinese utterance, the Moses toolkit [14] was used to train an SMT system on all clean text to generate 20 translation hypotheses for each ASR hypothesis. The best translation (among 20x20=400 hypotheses) is determined by the highest translation score output by the Moses decoder.

### 4.3. The Proposed System

The proposed speech translation system differs from the baseline system in two important aspects. First, for each Chinese utterance with 400 translation hypotheses (20 from ASR with each producing 20 SMT choices), the best translation candidate is determined by the highest integrative score using Eq.2 instead of only using the translation score. Second, the proposed decision-feedback training was applied to optimize parameters in the Chinese LM, English LM and phrase translation table. The parameters in the decision-feedback training were tuned on the 296 development set. The 294 test set was used to evaluate the speech translation performance.

### 4.4. Results

In Fig.1, the red line illustrates the BLEU score of the training set against the number of the MCE iterations, and the straight blue line is the BLEU score of the baseline system. One interesting point to note is that at iteration zero, the new system already outperforms the baseline system by 0.65 points. This indicates that the use of the

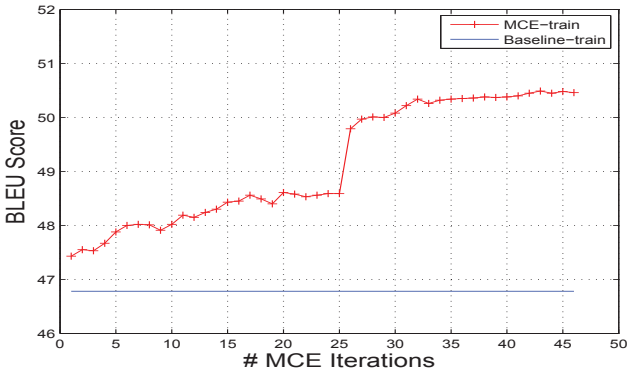


**Fig. 1**. *BLEU scores on the training set over the MCE iterations*

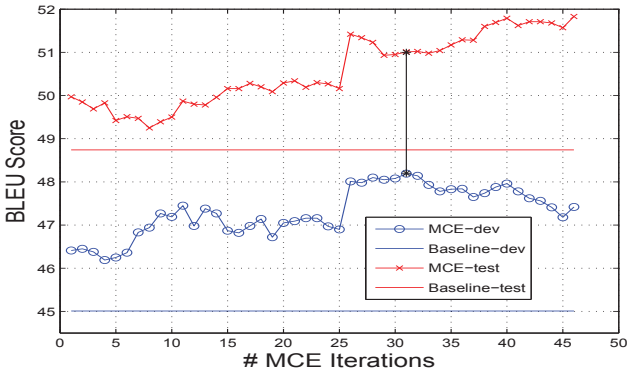


**Fig. 2**. *BLEU scores on the development and test sets over the MCE iterations*

integrative decision function alone can improve the translation performance without any help of discriminative training. The reason for this benefit is that the integrative decision function provides an end-to-end measure of the translation quality. For example, an English translation hypothesis with a low score in the baseline system may be promoted to be the best candidate if its corresponding ASR hypothesis has a high acoustic score and/or Chinese LM score. On the other hand, the best English translation candidate in the baseline system may be degraded if its corresponding ASR score and/or the Chinese LM score is poor. Another point to note is that a sharp increase in BLEU (around 1.5 points) happened between $24^{th}$ and $26^{th}$ iteration. By carefully examining the translation output in these iterations, we found that this sharp change is caused by the discontinuity in the BLEU score measurement. For example, in one case, the best translation hypothesis has a BLEU score of 22.7 while the second best hypothesis is 30.3. After several rounds of the MCE training, if the second best hypothesis manages to beat the original best translation, the BLEU score would have a significant change.

Fig.2 illustrates the translation performance on the development set (blue) and test set (red). The baseline performance for the development and test set are in straight lines. Based on the BLEU score on the development set, we selected the $31^{th}$ iteration (denoted by the vertical line in black) for evaluating the translation performance on the test set. Comparing with the baseline, the proposed system improves the BLEU score by 2.3 points. The improvement is particularly encouraging given that only a small amount (2,362 utterances) of parallel data with acoustic scores was available to apply the decision-feedback training, and the SMT system was able to quickly adapt to accept sentences with ASR errors.

## 5. CONCLUSION AND FUTURE WORK

In this paper a novel decision function for scoring the hypotheses of speech translation was presented. The associated decision-feedback training that embeds the decision function was used to further improve the translation performance in terms of the BLEU score. The experimental results demonstrated the effectiveness of both the decision function and the associated discriminative training algorithm.

The decision-feedback training framework presented in this paper not only permits the estimation of LM and phrase translation parameters as we have experimented so far, it can also be applied to estimate HMM parameters if the audio recordings of the training data are available. In our recent work [3], a log-linear model is used for speech translation, including not only "features" derived from Bayesian analysis but a number of other features contributing to the speech translation quality. It was found that with the end-to-end optimization on the feature weights, an improved BLUE score is obtained even at the cost of increased word error rate in speech recognition. We will also explore the use of an advanced optimization technique – growth-transform [2] to improve the current gradient descent optimization method. The growth-transform technique allows the removal of the approximations made in deriving the decision function in Eq.2 so that richer information for competing candidates in ASR can be exploited in the end-to-end optimization for speech translation.

## 6. REFERENCES

[1] A. Bozarov, Y. Sagisaka, R. Zhang and G. Kikui, "Improved speech recognition word lattice translation by confidence measure," in *Proc. Interspeech*, 2005, pp. 3197–3200.

[2] X. He, L. Deng and C. Wu, "Discriminative learning in sequential pattern recognition – A unifying review for optimization-oriented speech recognition," *IEEE Signal Processing Magazine*, vol. 25, no. 5, pp. 14-36, 2008.

[3] X. He, L. Deng and A. Acero, "Why is word error rate not a good metric for speech recognizer training for the speech translation task?", submitted to ICASSP 2011.

[4] B.-H. Juang, W. Hou and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Trans. SAP*, vol. 5, no. 3, pp. 257 –265, 1997.

[5] P. Koehn, F. J. Och and D. Marcu, "Statistical phrase-based translation," in *Proc. NAACL*, 2003, pp. 48–54.

[6] E. Matusov, S. Kanthak and H. Ney, "Integrating speech recognition and machine translation: Where do we stand?" in *Proc. ICASSP*, 2006.

[7] S. Matsoukas, I. Bulyko, B. Xiang, K. Nguyen, R. Schwartz and J. Makhoul, "Integrating speech recognition and machine translation," in *Proc. ICASSP*, 2007, pp. 1281–1284.

[8] P. Liang, A. Bouchard-Cote, D. Klein and B. Taskar, "An end-to-end discriminative approach to machine translation," in *Proc. COLING-ACL*, 2006.

[9] K. Papineni, S. Roukos, T. Ward and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. ACL*, 2002.

[10] H. Ney, "Speech translation: Coupling of recognition and translation," in *Proc. ICASSP*, 1999, pp. 1149–1152.

[11] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proc. ACL*, 2003, pp. 160–167.

[12] S. Saleem, S. Jou, S. Vogel and T. Schultz, "Using word lattice information for a tighter coupling in speech translation systems," in *Proc. ICSLP*, 2004, pp. 41–44.

[13] S. Yaman, L. Deng, D. Yu, Y.-Y. Wang and A. Acero, "An integrative and discriminative technique for spoken utterance classification," *IEEE Trans. ASLP*, vol. 16, no. 6, pp. 1207–1214, 2008.

[14] The Moses toolkit, "`http://www.statmt.org/moses/`".