Xiaodong He and Li Deng

Speech Recognition, Machine Translation, and Speech Translation—A Unified Discriminative Learning Paradigm

In the past two decades, significant progress has been made in automatic speech recognition (ASR) [2], [9] and statistical machine translation (MT) [12]. Despite some conspicuous differences, many problems in ASR and MT are closely related and techniques in the two fields can be successfully cross-pollinated. In this lecture note, we elaborate on the fundamental connections between ASR and MT, and show that the unified ASR discriminative training paradigm recently developed and presented in [7] can be extended to train MT models in the same spirit.

In addition, inspired by substantial success in ASR and MT, speech translation (ST), which aims to translate speech sound from one language to another and can be regarded as ASR and MT in tandem, has recently attracted increasing attention among speech and natural language communities. In the article published in a recent issue of *IEEE Signal Processing Magazine*, Treichler [25] included "universal translation" as one of the future needs and wants that are enabled by signal processing technology. In fact, a full range of research has been taking place today on all components of such universal translation, not only with text but also with speech [21], [18]. However, compared with ASR or MT, ST presents particularly challenging problems. Unlike individual ASR or MT systems, an ST system usually consists of both ASR and MT as subcomponents, in which ASR and MT are tightly coupled and interact with each other. Therefore, both the complexity and capability of such a complex system increase dramatically and effective model learning for the system becomes particularly crucial. The effective discriminative training framework [7] developed for ASR is shown to extend naturally not only to MT but also to ST.

Digital Object Identifier 10.1109/MSP.2011.941852
Date of publication:

INTRODUCTIONS TO ASR, MT, AND ST

ASR is the process and the related technology for converting a speech signal into a sequence of words (or other linguistic symbols). Modern general-purpose speech recognition systems are usually based on hidden Markov models (HMMs) for acoustic modeling and N-gram Markov models for language modeling. Over the past several decades, ASR has achieved substantial progress in four major areas: First, in the infrastructure area, Moore's law, in conjunction with the constantly shrinking cost of memory, has been instrumental in enabling speech recognition researchers to develop and run increasingly complex systems. The availability of common speech corpora for speech system training, development, and evaluation has been critical in creating systems of increasing capabilities. Second, in the area of knowledge representation, major advances in speech signal representations have included perceptually motivated acoustic features of speech. Third, in the area of modeling and algorithms, the most significant paradigm shift has been the introduction of statistical methods, especially of the HMM and extended methods for acoustic modeling. Fourth, in the area of recognition hypothesis search, key decoding or search strategies, originally developed in nonspeech applications, have focused on stack decoding (A* search), Viterbi, *N*-best, and lattice search/decoding.

MT refers to the process for converting text in one language (source) to its translated version in another language (target). The history of machine translation can be traced back to the 1950s [10]. In the early 1990s, Brown et al. conducted the pioneering work in introducing a statistical modeling approach to MT and in establishing a range of IBM MT models—Models 1–5 [3]. Since then, a series of important progress has been accomplished in a full span of the MT component technologies including word alignment [17], phrase-based MT methods [11], hierarchical phrase-based methods [4], syntax-based MT methods [5], discriminative training MT methods [13], [16], and system combination methods [23]. Today, MT has been in wide public use, e.g., via the Google translation service (http://translate.google.com) and Microsoft translation service (http://translator.bing.com).

An ST system takes the source speech signal as input and translates it into the target language, which is often in the format of text (but can also be in the format of speech through speech synthesis). ST usually consists of two major components: ASR and MT. Over the past several years, proposals have been made for the integration of these two components in an end-to-end ST task [15], [28]. In [15], Ney proposed a Bayes' rule-based integration of ASR and MT, in which the ASR output is treated as a hidden variable and the joint probability of the source speech input and the target text is modeled. Later, this approach was extended to a log-linear model by which the posterior probability of the translated text given input speech is modeled, where the feature functions are derived from the overall scores from probabilistic models such as speech

[TABLE 1] SUMMARY OF MAJOR BENCHMARK TASKS IN ASR, MT, AND ST.

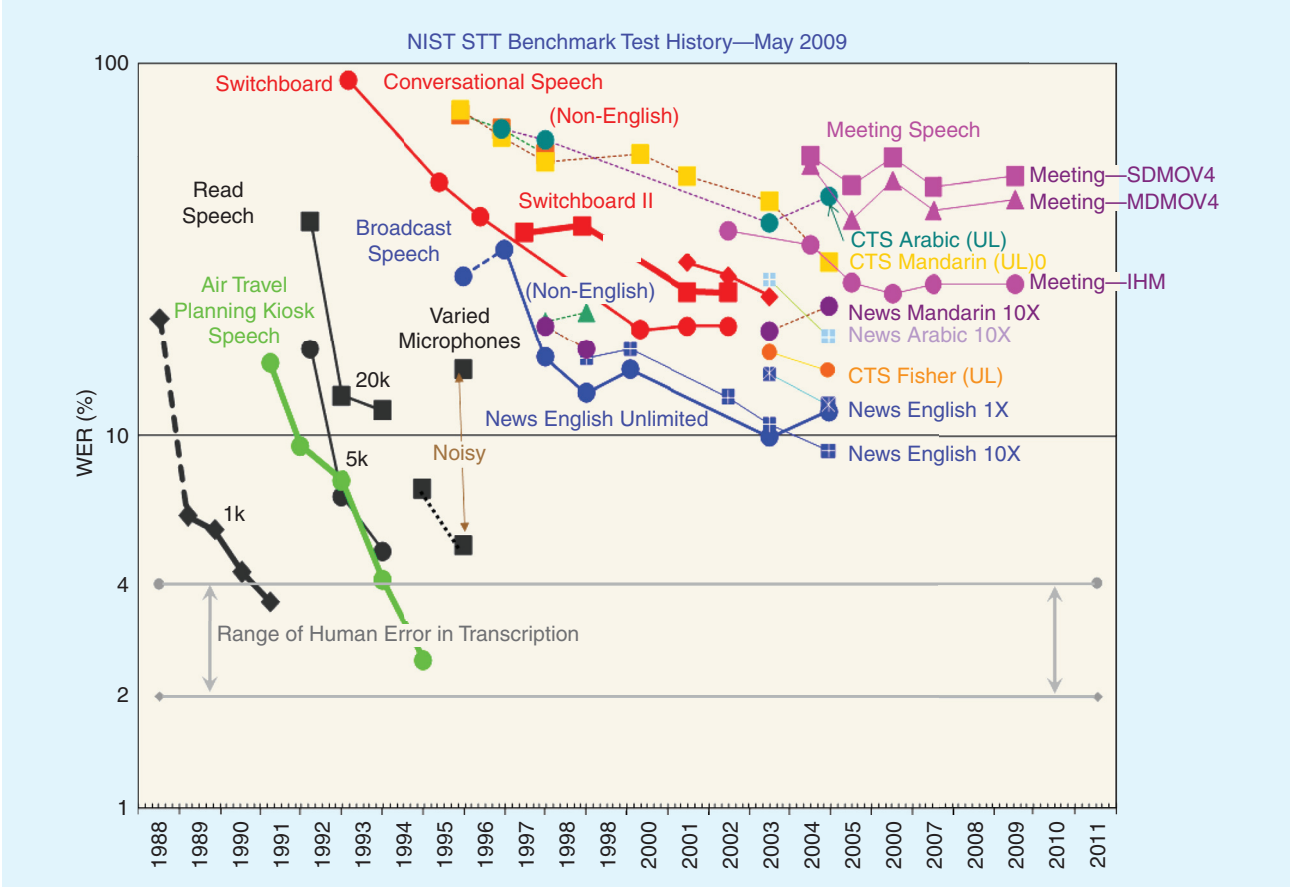
TASK NAME	CATEGORY	SCOPE AND SCENARIOS
TIMIT	ASR	English phonetic recognition; small vocabulary
WSJ 0&1	ASR	Mid-to-large vocabulary; dictation speech
AURORA	ASR	Small vocabulary, under noisy acoustic environments
DARPA EARS	ASR	Large vocabulary, broadcast, and conversational speech
NIST MT OPEN EVAL	MT	Large scale MT, newswire, and Web text
WMT (EUROPARL/EUROMATRIX)	MT	Large scale MT, newswire, and political text
C-STAR	ST	Limited domain spontaneous speech
DARPA TRANSTAC	ST	Limited domain spontaneous dialog
IWSLT	ST	Limited domain dialog and unlimited free-style talk
TC-STAR	ST	Broadcast speech, political speech
DARPA GALE	ST	Broadcast speech, conversational speech

acoustic model, source language model, probabilistic MT model, and target language model [14], [27]. The integration of ASR and MT in ST has been proposed and implemented mainly at the decoding stage. Little work was done on optimal integration for the construction or learning of ST systems.

On the other hand, discriminative learning has been used pervasively in recent statistical signal processing and pattern recognition research including ASR and MT separately [7], [13], [26]. In particular, a unified discriminative training framework was proposed for ASR, in which a variety of discriminative training criteria are consolidated into one single general mathematical form, and a general optimization technique based on growth transformation was developed [7]. Applications of this optimization framework to MT and the extension to ST will be presented in this lecture note.

To conclude this brief introduction, we point out that in each of the ASR, MT, and ST fields, a series of benchmark evaluations have been conducted, historically and some continuing until this as well as coming years, with active participation of international research teams. In Table 1, we summarized major benchmark tasks for each of the three fields.

Historically, NIST have conducted a comprehensive series of tests on ASR over two and a half decades, and the performance of major ASR benchmark tests is summarized in [19]. Here we reproduce the results in Figure 1 for the readers' reference.



[FIG1] After [19], a summarization of historical performance progresses on major ASR benchmark tests conducted by NIST.

SR AND MT—CLOSELY RELATED PROBLEMS AND TECHNICAL APPROACHES

In this section, we present a tutorial on technical aspects of ASR and MT. We will focus on the similarity of the ASR and MT processes.

Current state-of-the-art ASR systems are mainly based on HMM for acoustic modeling. Generally, it is assumed that the speech signal and its features are a realization of some semantic or linguist message encoded as a sequence of linguistic symbols. To recognize the underlying symbol sequence given a spoken utterance, the speech waveform is first converted into a sequence of feature vectors equally spaced in time. The feature extraction process is illustrated in Figure 2. Each feature vector is assumed to represent the speech waveform over a short duration of about 10 ms, within which the speech waveform is regarded as a stationary signal. Typical parametric representations include linear prediction coefficients, perceptual linear prediction, and Mel frequency cepstral coefficients (MFCCs), plus their time derivatives. These vectors are usually considered as independent observations given a state of the HMM. In the cases where discrete HMM is used for acoustic modeling, vector quantization is further applied to label each feature vector with a discrete symbol.

Given the speech signal being represented by a sequence of observation vectors X, the ASR process is illustrated in Figure 3.

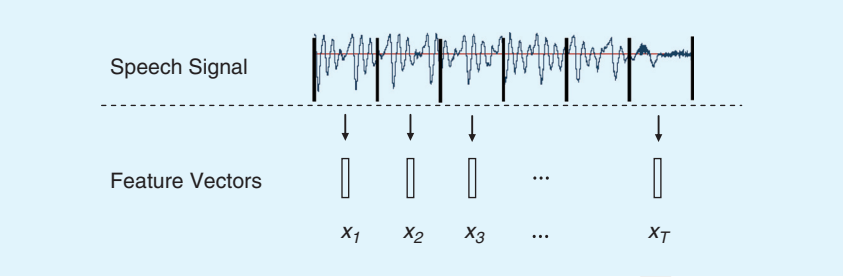
In ASR, the optimal word sequence F given X is obtained by

$$\begin{aligned}\hat{F} &= \arg\max_F P(F|X, \Lambda)\\ &= \arg\max_F P(F, X|\Lambda)\\ &= \arg\max_F P(X|F, \Lambda)P(F|\Lambda), \end{aligned} \quad (1)$$

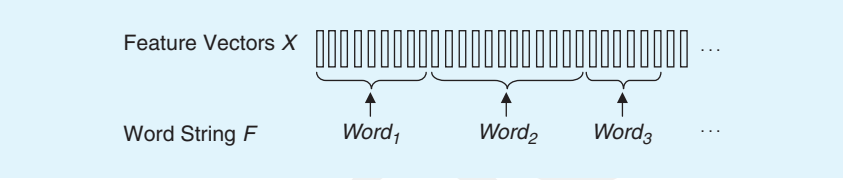
where $p(X|F, \Lambda)$ is often represented by an HMM [22], parameterized by Λ; i.e.,

$$\begin{aligned}&p(X|F, \Lambda)\\ &= \sum_q p(q|F, \Lambda)\cdot p(X|q, \Lambda)\\ &= \sum_q \prod_t p(q_t|q_{t-1}, \Lambda)\prod_t p(x_t|q_t, \Lambda), \end{aligned} \quad (2)$$

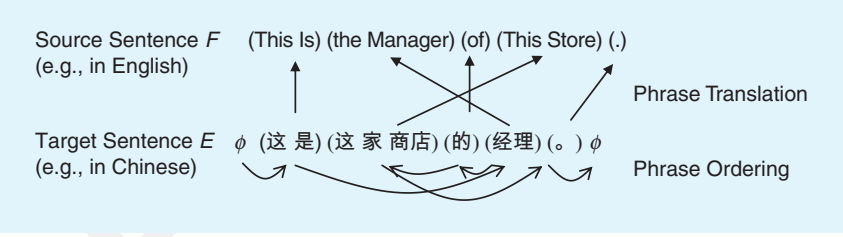
where q is the HMM state sequence and



[FIG2] **Illustration of the speech feature vector extraction process. The raw waveform of speech is parameterized to a sequence of observation vectors.**



[FIG3] **Illustration of the speech recognition process. The word string that corresponds to that observation sequence is decoded by the recognizer.**



[FIG4] **Illustration of the phrase based machine translation process. Note the nonmonotonic phrase ordering between the source sentence and the target sentence. This distinguishes the decoding process of MT from that of ASR.**

$p(X|q, \Lambda)$ and $p(q|F, \Lambda)$ are determined by the emitting probability and transition probability of the HMM, respectively. (Here we use F, instead of more commonly used W, to denote the word sequence to facilitate the comparison between ASR and MT, which is to be discussed shortly.)

In (1), $P(F|\Lambda)$ is usually modeled by an N-gram language model, i.e., an $(N-1)$th order Markov chain

$$p(F|\Lambda) = \prod_i^L p(f_i|f_{i-1}, \ldots, f_{i-N+1}, \Lambda), \quad (3)$$

where f_i is the ith word in the sentence F.

Similar to ASR, an MT system takes a sequence of observation symbols, e.g., words in the source language, and converts it into another sequence of symbols, e.g., words in the target language. A simple phrase-based MT system is illustrated in Figure 4.

Take the phrase-based MT paradigm as an example. Following Bayes' rule, the optimal translation E (English, often used as the target language) of the source (or foreign) language sentence F is

$$\begin{aligned}\hat{E} &= \arg\max_E P(E|F, \Lambda)\\ &= \arg\max_E P(E, F|\Lambda)\\ &= \arg\max_E P(F|E, \Lambda)P(E|\Lambda) \end{aligned} \quad (4)$$

In MT, $p(F|E, \Lambda)$ can be modeled by a phrase-based translation model, e.g.,

$$\begin{aligned}&p(F|E, \Lambda)\\ &= \sum_a p(a|E, \Lambda)\cdot p(F|E, a, \Lambda)\\ &= \sum_a \prod_l p(a_l|a_{l-1}, \Lambda)\prod_l p(f_l|e_l, \Lambda), \end{aligned} \quad (5)$$

where a is the phrase sequence, which may not be monotonic from left to right as the state sequence for HMM, as determined by the segmentation and order-

[TABLE 2] AFTER TABLE 1 OF [7], BY TAKING DIFFERENT FORM OF THE CLASSIFICATION QUALITY MEASURE $C_{DT}(F_1 \ldots F_R)$, THE UNIFIED CRITERION COVERS DIFFERENT POPULAR DISCRIMINATIVE TRAINING CRITERIA SUCH AS MMI, MCE, MPE/MWE. F_r^* IS THE REFERENCE OF THE rth SENTENCE.

ASR METRIC	$C_{DT}(F_1 \ldots F_R)$
MMI	$\prod_r \delta(F_r, F_r^*)$
MCE	$\sum_r \delta(F_r, F_r^*)$
MPE/MWE	$\sum_r A(F_r, F_r^*)^\dagger$

$^\dagger A(F_r, F_r^*)$ is the raw phone/word accuracy count of F_r given the transcription reference F_r^*.

ing model. In a conventional phrase-based MT system, a uniform distribution is used for segmentation and the probability of jumping from one phrase to another is just a function of the jumping distance. On the other hand, $p(F|E, a, \Lambda) = \prod_l p(f_l|e_l, \Lambda)$ are computed by the phrase translation model, where f_l and e_l are the lth phrase in the source sentence and the target sentence (after reordering), respectively [11].

Comparing the ASR process in Figure 3 and the MT process in Figure 4, we see a striking similarity between the two, with only one major difference that MT includes a nonmonotonic decoding process while the order of the output symbols of ASR is monotonic to the order of the input symbols. Importantly, the HMM-based acoustic model in ASR [see (2)] and the phrase-based translation model in MT with fixed reordering [see (5)] have essentially the same mathematical form. In the next section, we will show that this similarity makes it possible to extend the unified discriminative training approach already developed for ASR [7] to discriminative MT training.

A UNIFIED LEARNING PARADIGM FOR ASR AND MT

In [7], a general parameter learning criterion was presented for ASR. In this lecture note, we show that it can be naturally extended to MT (this section) and further to ST (see the section "Extending the Unified Learning Paradigm to Speech Translation") based on the analysis provided in the preceding section.

Let us first briefly review the growth-transformation (GT)-based discriminative training approach to ASR. The unified discriminative training criterion proposed in [7] takes the form shown in (6) at the bottom of the page, where R is the number of sentences in the training set. We denote by Λ the parameter set, and denote by X_i and F_i the speech feature sequence and the recognition symbol sequence of ith utterance, respectively. $C_{DT}(F_1 \ldots F_R)$ is a classification quality measure independent of Λ. As shown in Table 2, by taking different forms of $C_{DT}(F_1 \ldots F_R)$, this general objective function covers different common discriminative training criteria such as maximum mutual information (MMI), minimum classification error (MCE), and minimum phone/word error (MPE/MWE).

As demonstrated in [7], the general objective function can be optimized iteratively through the construction of a series of GT; i.e., given (6), if we denote by $G(\Lambda)$ the numerator of $O(\Lambda)$ and $H(\Lambda)$ the denominator of $O(\Lambda)$, a primary auxiliary function can be constructed taking the form of

$$F(\Lambda; \Lambda') = G(\Lambda) - O(\Lambda')H(\Lambda) + D, \quad (7)$$

where Λ' is the parameter set obtained through the last immediate iteration, and D is a Λ independent quantity. It is easy to see that a growth of the value $F(\Lambda; \Lambda')$ will guarantee a growth of the value of $O(\Lambda)$.

In the following, we denote by $\boldsymbol{X} = X_1 \ldots X_R$ the super-string of speech features, i.e., the concatenation of all speech feature sequences or strings, and $\boldsymbol{F} = F_1 \ldots F_R$ the super-string of recognition symbols. Then, the primary auxiliary function takes the following form:

$$\begin{aligned} F(\Lambda; \Lambda') &= \sum_{\boldsymbol{F}} p(\boldsymbol{X}, \boldsymbol{F}|\Lambda) C(\boldsymbol{F}) \\ &\quad - O(\Lambda') \sum_{\boldsymbol{F}} p(\boldsymbol{X}, \boldsymbol{F}|\Lambda) + D \\ &= \sum_{\boldsymbol{F}} \sum_{\boldsymbol{q}} \sum_{\boldsymbol{\chi}} [\Gamma(\Lambda') + D] \\ &\quad \times p(\boldsymbol{\chi}, \boldsymbol{F}, \boldsymbol{q}|\Lambda), \end{aligned} \quad (8)$$

where $\Gamma(\Lambda') = \delta(\boldsymbol{X}, \boldsymbol{\chi})[C(\boldsymbol{F}) - O(\Lambda')]$ is a Λ-independent quantity, and we denote by $\boldsymbol{q}$ the super-string of HMM state sequences. $\boldsymbol{\chi}$ is a super-string in the space of $\boldsymbol{X}$.

We then further construct the second auxiliary function as shown in [7], [1], and [6], i.e.,

$$\begin{aligned} &V(\Lambda; \Lambda') \\ &= \sum_{\boldsymbol{F}} \sum_{\boldsymbol{q}} \sum_{\boldsymbol{\chi}} [\Gamma(\Lambda') + D] \\ &\quad \times p(\boldsymbol{\chi}, \boldsymbol{F}, \boldsymbol{q}|\Lambda') \cdot \log p(\boldsymbol{\chi}, \boldsymbol{F}, \boldsymbol{q}|\Lambda). \end{aligned} \quad (9)$$

It can be shown that an increase in the auxiliary function $V(\Lambda; \Lambda')$ guarantees an increase in $F(\Lambda; \Lambda')$ when the constant D is sufficiently large. Importantly, $V(\Lambda; \Lambda')$ is easier to optimize than $F(\Lambda; \Lambda')$, because the new logarithm $\log p(\boldsymbol{\chi}, \boldsymbol{F}, \boldsymbol{q}|\Lambda)$ introduced in (9) can lead to significant simplification of $V(\Lambda; \Lambda')$ using the Markovian property of the HMM.

In the following, we briefly show the procedure of optimizing $V(\Lambda; \Lambda')$ w.r.t. to the parameters of the acoustic model of speech in ASR, using the emitting probability of the HMM as an example.

The terms of $V(\Lambda; \Lambda')$ that are relevant to the optimization of emitting probability of the HMM can form an equivalent auxiliary function

$$\begin{aligned} &U(\Lambda; \Lambda') \\ &= \sum_{\boldsymbol{F}} \sum_{\boldsymbol{q}} [C(\boldsymbol{F}) - O(\Lambda')] \end{aligned}$$

$$O(\Lambda) = \frac{\sum_{F_1 \ldots F_R} p(X_1 \ldots X_R, F_1 \ldots F_R|\Lambda) \cdot C_{DT}(F_1 \ldots F_R)}{\sum_{F_1 \ldots F_R} p(X_1 \ldots X_R, F_1 \ldots F_R|\Lambda)} \quad (6)$$

$$\times p(\boldsymbol{X},\boldsymbol{F},\boldsymbol{q}|\Lambda')\cdot\log p(\boldsymbol{X}|\boldsymbol{q},\Lambda) + D\sum_{\boldsymbol{F}}\sum_{\boldsymbol{q}}\sum_{\boldsymbol{\chi}} p(\boldsymbol{\chi},\boldsymbol{F},\boldsymbol{q}|\Lambda') \cdot \log p(\boldsymbol{\chi}|\boldsymbol{q},\Lambda). \quad (10)$$

In the discrete HMM case, the emitting probability is a multinomial distribution, i.e.,

$$b_i(k) = p(x_t = k|q_t = i, \Lambda), \quad (11)$$

where $k = 1, 2, \ldots, K$ is the index of the vector quantization codebook, and $\sum_{k=1}^{K} b_i(k) = 1$.

After applying the Lagrange multiplier method to maximize $U(\Lambda; \Lambda')$ w.r.t. the emitting probability distribution, we can obtain the following re-estimation formula for $\{b_i(k)\}$:

$$b_i(k) = \frac{\sum_{t=1}^{T}\sum_{s.t.\,x_t=k}\Delta\gamma(i,t) + D_i b_i'(k)}{\sum_{t=1}^{T}\Delta\gamma(i,t) + D_i}, \quad (12)$$

where D_i is a Λ-independent quantity [7], and

$$\Delta\gamma(i,t) = \sum_{\boldsymbol{F}} p(\boldsymbol{F}|\boldsymbol{X},\Lambda')[C(\boldsymbol{F}) - O(\Lambda')]\gamma_{i,\boldsymbol{F}}(t) \quad (13)$$

and

$$\gamma_{i,\boldsymbol{F}}(t) = p(q_t = i|\boldsymbol{X},\boldsymbol{F},\Lambda') = \sum_{\boldsymbol{q}:q_t=i} p(\boldsymbol{q}|\boldsymbol{F},\boldsymbol{X},\Lambda') \quad (14)$$

is the occupation probability of state i at time t.

Similarly, in the Gaussian density HMM case

$$b_i(k) \propto \frac{1}{|\Sigma_i|^{1/2}} \times \exp\left[-\frac{1}{2}(x_t-\mu_i)^T\Sigma_i^{-1}(x_t-\mu_i)\right]. \quad (15)$$

The mean vector and covariance matrix, μ_i and Σ_i, can be solved [7]. Taking the estimation of the mean vector as an example,

$$\mu_i = \frac{\sum_{t=1}^{T}\Delta\gamma(i,t)x_t + D_i\mu_i'}{\sum_{t=1}^{T}\Delta\gamma(i,t) + D_i}. \quad (16)$$

As discussed in the section "SR and MT—Closely Related Problems and Technical Approaches," MT has a close relationship with ASR. In particular, given the very similar formulations of (2) and (5) for the HMM-based acoustic modeling and phrase-based translation modeling, respectively, it is straightforward to extend the unified discriminative training criterion to phrase translation model learning. Here we provide some detail in the remainder of this section.

[TABLE 3] BY TAKING DIFFERENT FORMS OF THE CLASSIFICATION QUALITY MEASURE, $C_{DT}(\boldsymbol{E})$, THE UNIFIED CRITERION CORRESPONDS TO OPTIMIZATION OF DIFFERENT MT METRICS. IN THE TABLE, E_r^* IS THE TRANSLATION REFERENCE OF THE rTH SOURCE SENTENCE F_r. WE DENOTE BY $\boldsymbol{E}^*$ THE SUPER-STRING OF THE TRANSLATION REFERENCES.

MT Metric	$C_{DT}(\boldsymbol{E})$
BLEU	$BLEU(\boldsymbol{E}, \boldsymbol{E}^*)$
Avg. sentence-level BLEU	$\sum_r BLEU(E_r, E_r^*)/R$
TAR (e.g., 1-TER)	$\sum_r \lvert E_r^*\rvert \cdot [1 - TER(E_r, E_r^*)]$
Sentence-level error rate	$\sum_r \delta(E_r, E_r^*)$
Conditional likelihood	$\prod_r \delta(E_r, E_r^*)^{\dagger}$

†The unified criterion becomes the conditional likelihood of the translation reference given such a classification quality measure.

For MT, as shown in (4) and (5), the observation symbol sequence is the word sequence in the source (or foreign) language, F, and the recognition output is the word sequence in the target (e.g., English) language, E. Assuming the training corpus consists of R parallel sentence pairs, e.g., $\{(F_1, E_1), \ldots, (F_R, E_R)\}$. We denote by $\boldsymbol{F} = F_1 \ldots F_R$ the super-string of the source language, or the concatenation of all source word sequences. Likewise, we denote by $\boldsymbol{E} = E_1 \ldots E_R$ the super-string of the target language, or the concatenation of all translation hypotheses in the target language. Then, the objective function of discriminative training for MT becomes

$$O(\Lambda) = \frac{\sum_{\boldsymbol{E}} p(\boldsymbol{F},\boldsymbol{E}|\Lambda)\cdot C_{DT}(\boldsymbol{E})}{\sum_{\boldsymbol{E}} p(\boldsymbol{F},\boldsymbol{E}|\Lambda)}. \quad (17)$$

This general objective function of (17) is the model-based expectation of the classification quality measure $C_{DT}(\boldsymbol{E})$. Accordingly, by taking different forms of $C_{DT}(\boldsymbol{E})$, we will obtain a variety of objective functions that correspond to maximizing the expected value of different metrics, such as the (approximated) Bilingual Evaluation Understudy (BLEU) **<AU: please check that BLEU is spelled out correctly>** score [20], Translation Accuracy Rate (TAR) (which is defined as the complementary of the Translation Edit Rate (TER) [24], i.e., TAR = 1 – TER), etc. Some classification quality measure candidates are listed in Table 3.

It is noteworthy that, if $C_{DT}(\boldsymbol{E})$ takes a special form of $\prod_r \delta(E_r, E_r^*)$, the objective function will become the conditional likelihood of the translation reference, i.e.,

$$O(\Lambda) = \frac{p(\boldsymbol{F},\boldsymbol{E}^*|\Lambda)}{\sum_{\boldsymbol{E}} p(\boldsymbol{F},\boldsymbol{E}|\Lambda)} = p(\boldsymbol{E}^*|\boldsymbol{F},\Lambda). \quad (18)$$

Given the virtually identical mathematical formulations of ASR and phrase-based MT, i.e., (2) and (6) versus (5) and (17), optimization of (17) naturally takes the same steps of derivation of that of (6). Using the phrase translation model as an example, the phrase translation probability follows a multinomial distribution, similar to the discrete HMM case just discussed. We denote by f_k the kth entry in the source language phrase list, and by e_i the ith entry in the target language phrase list, the translation probability of f_k given e_i can be written in the same form as the emission probability of a discrete HMM

$$b_i(k) = p(f_k|e_i). \quad (19)$$

Consequently, following (12), the GT re-estimation formula of $b_i(k)$ becomes

$$b_i(k) = \frac{\sum_{l=1}^{L} \Delta\gamma(i,l) + D_i b_i'(k)}{\sum_{\substack{l=1 \\ s.t. f_l = f_k}}^{L} \Delta\gamma(i,l) + D_i}, \tag{20}$$

where

$$\Delta\gamma(i,l) = \sum_{\boldsymbol{E}} p(\boldsymbol{E}|\boldsymbol{F}, \Lambda') \times [C(\boldsymbol{E}) - O(\Lambda')]\gamma_{i,\boldsymbol{E}}(l) \tag{21}$$

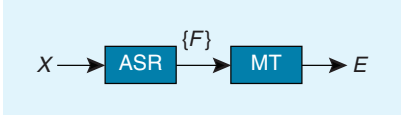
and

$$\gamma_{i,\boldsymbol{E}}(l) = p(e_l = e_i|\boldsymbol{E}, \boldsymbol{F}, \Lambda') = \sum_{\boldsymbol{a}: e_{a_l} = e_i} p(\boldsymbol{a}|\boldsymbol{E}, \boldsymbol{F}, \Lambda'). \tag{22}$$

Theoretically, one can take any arbitrary classification quality measure as the objective function to fit the above optimization framework. In practice, however, only a limited number of choices are feasible, making the computation of (21) tractable. Readers are referred to [7] and [8] for more detail on how (21) can be efficiently computed.

SPEECH TRANSLATION AS SR AND MT IN TANDEM

ST takes the source speech signal as input and produces as output the translated text of that utterance in another language. A general architecture for ST is illustrated in Figure 5. The input speech signal X is first fed into an ASR component, and the ASR component generates the recognition output set $\{F\}$, which is in the source language. The recognition hypothesis set $\{F\}$ is finally passed to the MT component to obtain the translation sentence E in the target language. In the following, we use F to represent an ASR hypothesis in the N-best list, while in practice a lattice is usually used as a compact representation of a N-best list consisting of many ASR hypotheses.



[FIG5] Two components in tandem in a full speech translation system

Let us denote by Λ the aggregation of all model parameters of a full ST system to be optimized. With fixed Λ, the optimal translation $\hat{E}$ given input speech X is obtained via the decoding process according to

$$\hat{E} = \arg\max_{E} P(E|X, \Lambda) = \arg\max_{E} P(E, X|\Lambda). \tag{23}$$

In this note, we follow the Bayes' rule-based integration of ASR and MT proposed in [15]. For each sentence, we decompose the joint probability between speech signal X and its translation E into the following form:

$$\begin{aligned} p(X, E|\Lambda) &= \sum_{F} p(X, F, E|\Lambda) \\ &= \sum_{F} p(E|\Lambda) p(F|E, \Lambda) \times p(X|F, E, \Lambda) \\ &\approx \sum_{F} p(E|\Lambda) p(F|E, \Lambda) \times p(X|F, \Lambda), \end{aligned} \tag{24}$$

where F is a recognition hypothesis of X in the source language, $p(X|F, \Lambda)$ is the likelihood of the speech signal given a transcript (e.g., the recognition score), and $p(F|E, \Lambda)$ is the likelihood of the source sentence given the target sentence, (e.g., the translation score). And $p(E|\Lambda)$ is the LM score of the target sentence E.

Conventionally, all these ASR, translation, and language models are trained separately and are optimized by different criteria. In the next section, we extend the GT-based general discriminative training framework above to train all these model parameters jointly so as to optimize the end-to-end ST quality.

EXTENDING THE UNIFIED LEARNING PARADIGM TO SPEECH TRANSLATION

After reusing the same definitions of super-strings $\boldsymbol{X}$, $\boldsymbol{F}$, $\boldsymbol{E}$, as provided earlier, the general discriminative training objective function for ST becomes

$$O(\Lambda) = \frac{\sum_{\boldsymbol{E},\boldsymbol{F}} p(\boldsymbol{X}, \boldsymbol{F}, \boldsymbol{E}|\Lambda) \cdot C_{DT}(\boldsymbol{E})}{\sum_{\boldsymbol{E},\boldsymbol{F}} p(\boldsymbol{X}, \boldsymbol{F}, \boldsymbol{E}|\Lambda)}. \tag{25}$$

Similar to (6) for ASR and to (17) for MT, the general objective function of (25) is model-based expectation of the classification quality measure $C_{DT}(\boldsymbol{E})$ for ST. Accordingly, by taking different forms of $C_{DT}(\boldsymbol{E})$, we will obtain a variety of objective functions that correspond to maximizing the expected value of different ST evaluation measures or metrics. For ST, the classification quality metrics are similar to those listed in Table 3 for MT.

However, optimization for (25) is more complicated than that for (6) or (17). To derive GT re-estimation formulas for the models of a ST system, we can construct a somewhat more complex primary auxiliary function of

$$\begin{aligned} F(\Lambda; \Lambda') &= \sum_{\boldsymbol{E}} \sum_{\boldsymbol{F}} p(\boldsymbol{X}, \boldsymbol{F}, \boldsymbol{E}|\Lambda) C(\boldsymbol{E}) \\ &\quad - O(\Lambda') \sum_{\boldsymbol{E}} \sum_{\boldsymbol{F}} p(\boldsymbol{X}, \boldsymbol{F}, \boldsymbol{E}|\Lambda) + D \\ &= \sum_{\boldsymbol{E}} \sum_{\boldsymbol{F}} \sum_{\boldsymbol{a}} \sum_{\boldsymbol{q}} \sum_{\boldsymbol{\chi}} [\Gamma(\Lambda') \\ &\quad + D] p(\boldsymbol{\chi}, \boldsymbol{F}, \boldsymbol{E}, \boldsymbol{q}, \boldsymbol{a}|\Lambda), \end{aligned} \tag{26}$$

where $\Gamma(\Lambda') = \delta(\boldsymbol{X}, \boldsymbol{\chi})[C(\boldsymbol{E}) - O(\Lambda')]$ is a Λ-independent quantity. We denote by $\boldsymbol{q}$ the super-string of HMM state sequences, and $\boldsymbol{a}$ the super-string of phrase segmentation and ordering sequences for the translation from $\boldsymbol{F}$ to $\boldsymbol{E}$. $\boldsymbol{\chi}$ is a super-string in the space of $\boldsymbol{X}$.

Similar to the previous derivations for ASR and MT, we also construct the secondary auxiliary function of

$$\begin{aligned} V(\Lambda; \Lambda') &= \sum_{\boldsymbol{E}} \sum_{\boldsymbol{F}} \sum_{\boldsymbol{a}} \sum_{\boldsymbol{q}} \sum_{\boldsymbol{\chi}} [\Gamma(\Lambda') + D] \\ &\quad \times p(\boldsymbol{\chi}, \boldsymbol{F}, \boldsymbol{E}, \boldsymbol{q}, \boldsymbol{a}|\Lambda') \\ &\quad \times \log p(\boldsymbol{\chi}, \boldsymbol{F}, \boldsymbol{E}, \boldsymbol{q}, \boldsymbol{a}|\Lambda'). \end{aligned} \tag{27}$$

In the following, we present the derivation and result of optimizing $V(\Lambda; \Lambda')$ w.r.t. to the parameters of the acoustic model of input speech, using Gaussian density HMM as an example, and to the parameters of the phrase translation probability distribution in the MT model.

The terms of $V(\Lambda;\Lambda')$ that are relevant to the optimization of the Gaussian density of the HMM form a new auxiliary function

$$\begin{aligned} U_{GM}(\Lambda;\Lambda') &= \sum_{\boldsymbol{E}}\sum_{\boldsymbol{F}}\sum_{\boldsymbol{q}}[C(\boldsymbol{E})-O(\Lambda')] \\ &\quad\times p(\boldsymbol{X},\boldsymbol{F},\boldsymbol{E},\boldsymbol{q}|\Lambda')\cdot\log p(\boldsymbol{X}|\boldsymbol{q},\Lambda) \\ &\quad + D\sum_{\boldsymbol{E}}\sum_{\boldsymbol{F}}\sum_{\boldsymbol{q}}\sum_{\boldsymbol{\chi}} p(\boldsymbol{\chi},\boldsymbol{F},\boldsymbol{E},\boldsymbol{q}|\Lambda') \\ &\quad\cdot\log p(\boldsymbol{\chi}|\boldsymbol{q},\Lambda). \end{aligned} \tag{28}$$

By setting $\partial U_{GM}(\Lambda;\Lambda')/\partial\mu_i = 0$ and $\partial U_{GM}(\Lambda;\Lambda')/\partial\Sigma_i = 0$, following similar derivation steps as in the section "SR and MT—Closely Related Problems and Technical Approaches," we can obtain the re-estimation formulas for all Gaussian parameters. Here we only present the derived formula for the mean vector as an example

$$\mu_i = \frac{\Sigma_t\,\Delta\gamma(i,t)\cdot x_t + D_i\mu_i'}{\Sigma_t\Delta\gamma(i,t)+D_i}, \tag{29}$$

where

$$\begin{aligned} \Delta\gamma(i,t) &= \sum_{\boldsymbol{E}}\sum_{\boldsymbol{F}} p(\boldsymbol{E},\boldsymbol{F}|\boldsymbol{X},\Lambda') \\ &\quad\times\{[C(\boldsymbol{E})-O(\Lambda')]\gamma_{i,\boldsymbol{F}}(t)\}. \end{aligned} \tag{30}$$

In (30), $\gamma_{i,\boldsymbol{F}}(t) = \sum_{\boldsymbol{q}:q_t=i} p(\boldsymbol{q}|\boldsymbol{F},\boldsymbol{X},\Lambda')$ is the conventional occupation probability of state i at time t, given the transcript $\boldsymbol{F}$ and the speech feature sequence $\boldsymbol{X}$.

Let us compare (30) for learning HMM parameters in ST with the counterpart of (13) in ASR. It is interesting to observe that, in ST, the core weight of the raw occupation probability $\gamma_{i,\boldsymbol{F}}(t)$, $[C(\boldsymbol{E})-O(\Lambda')]$, is based on the translation quality measure $C(\boldsymbol{E})$ at the target language, e.g., $C(\boldsymbol{E})$ could be the BLEU score of $\boldsymbol{E}$, and $O(\Lambda')$ is the model-based expected BLEU score. On the other hand, for ASR, the core weight of $\gamma_{i,\boldsymbol{F}}(t)$, i.e., $[C(\boldsymbol{F})-O(\Lambda')]$, is based on recognition quality measure $C(\boldsymbol{F})$ of the recognition hypothesis $\boldsymbol{F}$. Furthermore, in ASR, $\Delta\gamma(i,t)$ is the expectation of the weighted $\gamma_{i,\boldsymbol{F}}(t)$ over all recognition hypothesis $\{\boldsymbol{F}\}$, while in ST, $\Delta\gamma(i,t)$ is the expectation of the weighted $\gamma_{i,\boldsymbol{F}}(t)$ over all recognition hypotheses $\{\boldsymbol{F}\}$ and also over all the translation hypotheses $\{\boldsymbol{E}\}$.

The terms of $V(\Lambda;\Lambda')$ in (27) that are relevant to optimization of the phrase translation probability distribution give rise to another auxiliary function

$$\begin{aligned} U_{TM}(\Lambda;\Lambda') &= \sum_{\boldsymbol{E}}\sum_{\boldsymbol{F}}\sum_{\boldsymbol{a}}[C(\boldsymbol{E})-O(\Lambda')] \\ &\quad p(\boldsymbol{X},\boldsymbol{F},\boldsymbol{E},\boldsymbol{a}|\Lambda')\cdot\log p(\boldsymbol{F}|\boldsymbol{E},\boldsymbol{a},\Lambda) \\ &\quad + D\sum_{\boldsymbol{E}}\sum_{\boldsymbol{F}}\sum_{\boldsymbol{a}}\sum_{\boldsymbol{\chi}} p(\boldsymbol{\chi},\boldsymbol{F},\boldsymbol{E},\boldsymbol{a}|\Lambda') \\ &\quad\cdot\log p(\boldsymbol{F}|\boldsymbol{E},\boldsymbol{a},\Lambda). \end{aligned} \tag{31}$$

Applying the Lagrange multiplier and following similar derivation steps in the section "Speech Translation as SR and MT in Tandem," we obtain the re-estimation formula of

$$p(f_k|e_i) = b_i(k) = \frac{\sum_{\substack{l=1\\ s.t.f_l=f_k}}^{L}\Delta\gamma(i,l) + D_i b_i'(k)}{\sum_{l=1}^{L}\Delta\gamma(i,l)+D_i}, \tag{32}$$

where

$$\begin{aligned} \Delta\gamma(i,l) &= \sum_{\boldsymbol{E}}\sum_{\boldsymbol{F}} p(\boldsymbol{E},\boldsymbol{F}|\boldsymbol{X},\Lambda') \\ &\quad\times[C(\boldsymbol{E})-O(\Lambda')]\gamma_{i,\boldsymbol{E}}(l) \end{aligned} \tag{33}$$

and $\gamma_{i,\boldsymbol{E}}(l) = p(e_l=e_i|\boldsymbol{E},\boldsymbol{F},\Lambda') = \sum_{\boldsymbol{a}:e_{a_l}=e_i} p(\boldsymbol{a}|\boldsymbol{E},\boldsymbol{F},\Lambda')$.

Again, compared with the model estimation formula [see (21)] for MT discussed in the section "Speech Translation as SR and MT in Tandem," it is satisfying to observe that when the optimal ASR transcription is given, i.e., $\boldsymbol{F}$, were fixed at its optimal value instead of being treated as a (hidden) random variable, (33) would naturally reduce to (21). That is, the model learning problem for ST would reduce to that for a regular MT system.

SUMMARY AND DISCUSSION

ASR deals with numeric signals and MT deals with symbolic signals (i.e., text). While largely separate research communities have been working on ASR and MT, in this note we have shown that they can be formulated as two very similar optimization-oriented learning problems. The decoding problems for ASR and MT are also akin to each other but with one critical difference of nonmonotonicity for MT making the MT decoding problem more complex than the ASR counterpart. (See [9] and [12] for details of the decoding or search problems of ASR and MT, which are not covered in this lecture note.)

Due to the close relationship between ASR and MT, their respective technologies, the learning algorithms in particular, can be usefully cross-fertilized. The core of this lecture note did just that. Further, based on the insight provided by this cross-fertilization effort, we have been able to extend this learning framework to ASR and MT in tandem, which gives a principled way to design systems for speech ST-based on the same GT technique exploited throughout this note.

We have implemented the GT-based discriminative training method presented in this note in software, and successfully applied it in a number of experiments. Interested readers are referred to [7], [8], and [26] for the experimental results, as well as the techniques for approximations guided by the theoretical principle consistent with the rigorous solutions. Such principled corner cutting is a key to enabling successful applications in practice.

This lecture note provides an example of how a unified optimization framework can be usefully applied to solve a range of problems that may at a glance seem unrelated. It further shows how the same framework can be extended to construct a more complex system (e.g., ST) that consists of simpler subsystems (e.g., ASR and MT). Looking forward and generalizing further, we envision that even more complex intelligent systems that consist of multiple subsystems can be built in a similar way. When these subsystems, either in tandem or in parallel, are tightly incorporated and interacting with each other, an end-to-end discriminative learning scheme in the spirit of design philosophy as described in this note is likely to be especially

effective.

AUTHORS

Xiaodong He (xiaohe@microsoft.com) is a researcher and ***Li Deng*** (deng@microsoft.com) is a principal researcher, both at Microsoft Research, Redmond, Washington.

REFERENCES

[1] S. Axelrod, V. Goel, R. Gopinath, P. Olsen, and K. Visweswariah, "Discriminative estimation of subspace constrained Gaussian mixture models for speech recognition," *IEEE Trans. Audio Speech Lang. Processing*, vol. 15, no. 1, 2007, pp. 172–189.

[2] J. M. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, and D. O'Shaughnessy, "Research developments and directions in speech recognition and understanding," *IEEE Signal Processing Mag.*, vol. 26, pp. 75–80, May 2009.

[3] P. Brown, S. Pietra, V. Pietra, and R. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Comput. Linguistics*, vol. 19, 1994, pp. 263–311. **<AU: Please provide the issue number.>**

[4] D. Chiang, "Hierarchical phrase-based translation," *Comput. Linguistics*, vol. 33, pp. 201–228, 2007. **<AU: Please provide the issue number.>**

[5] M. Galley, J. Graehl, K. Knight, D. Marcu, S. DeNeefe, W. Wang, and I. Thayer, "Scalable inference and training of context-rich syntactic translation models," in *Proc. COLING-ACL*, 2006, pp. 961–968.

[6] A. Gunawardana and W. Byrne, "Discriminative speaker adaptation with conditional maximum likelihood linear regression," in *Proc. Eurospeech*, 2001, pp. 1203–1206.

[7] X. He, L. Deng, and W. Chou, "Discriminative learning in sequential pattern recognition," *IEEE Signal Processing Mag.*, vol. 25, no. 5, 2008, pp. 14–36.

[8] X. He and L. Deng, *Discriminative Learning for Speech Recognition: Theory and Practice*. San Rafael, CA: Morgan & Claypool Publishers, 2008.

[9] X. D. Huang and L. Deng, "An overview of modern speech recognition," in *Handbook of Natural Language Processing*, 2nd ed. London: Chapman & Hall, 2010, ch. 17. **<AU: Please provide the editor name(s) and page range.>**

[10] J. Hutchins, "From first conception to first demonstration: The nascent years of machine translation, 1947–1954," *Mach. Transl.*, vol. 12, 1997, pp. 195–252. **<AU: Please provide the issue number.>**

[11] P. Koehn, F. Och, and D. Marcu, "Statistical phrase-based translation," in *Proc. HLT-NAACL*, 2003. **<AU: Please provide the complete page range.>**

[12] P. Koehn, *Statistical Machine Translation*. Cambridge, U.K.: Cambridge Univ. Press, 2010.

[13] P. Liang, A. Bouchard-Cote, D. Klein, and B. Taskar, "An end-to-end discriminative approach to machine translation," in *Proc. of COLING-ACL*, 2006. **<AU: Please provide the complete page range.>**

[14] E. Matusov, S. Kanthak, and H. Ney, "Integrating speech recognition and machine translation: Where do we stand?" in *Proc. ICASSP*, 2006. **<AU: Please provide the complete page range.>**

[15] H. Ney, "Speech translation: Coupling of recognition and translation," in *Proc. ICASSP*, 1999. **<AU: Please provide the complete page range.>**

[16] F. Och, "Minimum error rate training in statistical machine translation," in *Proc. ACL*, 2003. **<AU: Please provide the complete page range.>**

[17] F. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Comput. Linguistics*, vol. 29, pp. 19–51, 2003. **<AU: Please provide the issue number.>**

[18] J. Olive. (2008). *Global autonomous language exploitation (GALE)*. [Online]. Available: http://www.darpa.mil/ipto/programs/gale/gale.asp

[19] D. Pallett, J. Fiscus, J. Garofolo, A. Martin, M. Przybocki, J. Ajot, M. Michel, W. Fisher, B. Lund, N. Dahlgren, and B. Tjaden. (2009). *The history of automatic speech recognition evaluations at NIST*. [Online]. Available: http://www.itl.nist.gov/iad/mig/publications/ASRhistory/index.html

[20] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. ACL*, 2002, pp. 311–318.

[21] M. Paul, M. Federico, and S. Stücker, "Overview of the IWSLT 2010 evaluation campaign," in *Proc. IWSLT*, Paris, France, 2010. **<AU: Please provide the complete page range.>**

[22] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, 1989. **<AU: Please provide the volume and issue numbers and complete page range.>**

[23] A. Rosti, B. Xiang, S. Matsoukas, R. Schwartz, N. Ayan, and B. Dorr, "Combining outputs from multiple machine translation systems," in *Proc. NAACL-HLT*, 2007, pp. 228–235.

[24] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proc. AMTA*, 2006. **<AU: Please provide the complete page range.>**

[25] J. Treichler, "Signal processing: A view of the future, Part 2," *IEEE Signal Processing Mag.*, vol. 26, pp. 83–86, May 2009.

[26] D. Yu, L. Deng, X. He, and A. Acero, "Large-margin minimum classification error training for large-scale speech recognition tasks," in *Proc. ICASSP*, 2008. **<AU: Please provide the complete page range.>**

[27] R. Zhang, G. Kikui, H. Yamamoto, T. Watanbe, F. Soong, and W.-K. Lo, "A unified approach in speech-to-speech translation: Integrating features of speech recognition and machine translation," in *Proc. COLING*, 2004. **<AU: Please provide the complete page range.>**

[28] Y. Zhang, L. Deng, X. He, and A. Acero, "A novel decision function and the associated decision-feedback learning for speech translation," in *Proc. ICASSP*, submitted for publication. **<AU: Please update the details. Has this been accepted?>**

[SP]