NOISE ROBUSTNESS IN SPEECH TO SPEECH TRANSLATION

Fu-Hua Liu, Yuqing Gao, Liang Gu, and Michael Picheny

IBM T.J. Watson Research Center
Yorktown Heights, NY 10598, U.S.A.
{fhl, yuqing, lianggu, picheny}@us.ibm.com

ABSTRACT

This paper describes various noise robustness issues in a speech-to-speech translation system. We present quantitative measures for noise robustness in the context of speech recognition accuracy and speech-to-speech translation performance. To enhance noise immunity, we explore two approaches to improve the overall speech-to-speech translation performance. First, a multi-style training technique is used to tackle the issue of environmental degradation at the acoustic model level. Second, a pre-processing technique, CDCN, is exploited to compensate for the acoustic distortion at the signal level. Further improvement can be obtained by combining both schemes. In addition to recognition accuracy for speech recognition, this paper studies and examines how closely speech recognition accuracy is related the overall speech-to-speech recognition. When we apply the proposed schemes to an English-to-Chinese translation task, the word error rate for our speech recognition subsystem is substantially reduced by 28% relative, to 13.2% from 18.9% for test data of 15dB SNR. The corresponding BLEU score improves to 0.478 from 0.43 for the overall speech-to-speech translation. Similar improvements are also observed for a lower SNR condition.

1. INTRODUCTION

The need to develop technologies to accomplish useful and satisfactory translation between languages is increasingly appreciated with rapid growth of internet applications and globalization of economy development [4,5,10,11]. With the introduction of speech technology into pervasive computing, the challenges for speech-to-speech translation demand different considerations from those for desktop applications. The adverse environment where these devices are usually deployed becomes very prevalent. The task becomes even more delicate when imperfect output of speech recognition is further used for machine translation.

Recently, we presented a speech translation system employing a statistical framework in a DARPA force protection domain[1]. Compared with another IBM system for an air travel domain[6], the force protection domain encompasses broader content coverage and, therefore, represents a more challenging task for translation.

Besides domain coverage, the issue of environmental degradation remains to be an important challenge. When speech recognition accuracy degrades harshly, the subsequent NLU and NLG will be affected severely as well. The extent of successful noise robustness results differs for speech recognition and speech-to-speech translation. To address the lack of correspondence between them, we attempt to evaluate our system improvement on two different metrics, word error rate for speech recognition and BLEU[2] for speech-to-translation when the proposed techniques are incorporated into system.

This paper is organized as follows, a brief overview of IBM's speech-to-speech translation system, MASTOR, is presented in Section2. Section 3 describes two efficient techniques to improve the noise immunity. Then, details of system setup, experiments and results will be given in Section 4. Finally, a conclusion and summary will be presented in Section5.

2. OVERVIEW OF SYSTEM

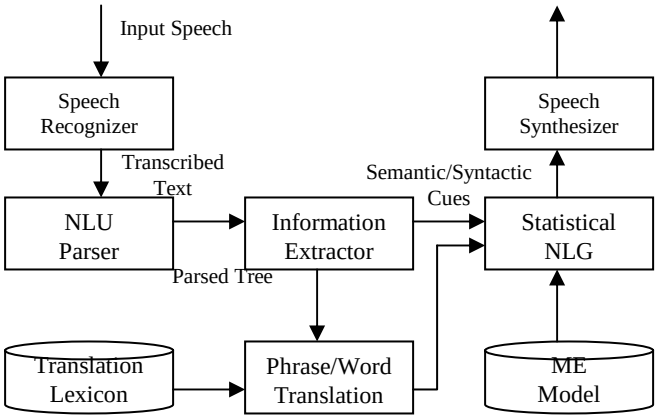

Figure 1: The architecture of MASTOR

MASTOR [1] (Multilingual Automatic Speech-To-Speech TranslatOR) is IBM's highly trainable speech-to-speech translation system, targeting conversational spoken language translation between English and Mandarin Chinese for limited domains. Figure 1 depicts the architecture of MASTOR. The speech input is processed and decoded by a large-vocabulary speech recognition system. Then the transcribed text is analyzed by a statistical parser [1,6] for semantic and syntactic features. A sentence-level natural language generator based on maximum entropy (ME) modeling [3] is used to generate sentences in the target language from the parser output. The produced sentence in target language is synthesized into speech by a high quality text-to-speech system [8].

The speech recognizer used in MASTOR is a large-vocabulary continuous speech recognition system. Both English and Mandarin speech recognizers share the same architecture with different acoustic models and language models. The recognition system is designed for a general-purpose dictation application with a powerful domain-independent trigram language model. In case of domain-specific applications, adapted N-gram language models can be derived handily from existing trigram models by interpolating with domain-specific language models. The output from speech recognition is furnished to the following NLU parser for further processing.

The NLU module [12] includes a statistical, decision-tree based parser. The parser extracts the semantic and syntactic information from input source sentence and produces a tree-structured semantic/syntactic representation, which is comparable to interlingua. The NLG system consists of a maximum entropy probability model using both the sentence level and concept level classes as constituents. The features used in the ME modeling include the previous symbols, local sentence or phrase type in the semantic tree, and the concept list that remains to be generated before current symbol. During the translation, a recursive search is performed on the parse tree of the input sentence in a bottom-up manner to generate the output word sequence in the target language.

.

3. IMPROVING NOISE IMMUNITY

Model adaptation techniques such as MLLR and MAP have been shown effective in reducing the mismatch between training and test condition. These approaches involve complex system re-configuration and expensive run-time requirement. Usually, they are accomplished during an off-line enrollment process. On the other hand, for real-time speech recognition in the speech-to-speech applications, the constantly varying ambience is a common cause for system performance degradation. The need to accommodate a varying degree of mismatch motivates us to focus on techniques that do not introduce hefty overhead in recognition.

3.1. Multi-Style Training

Multi-style training [13] has been shown to be a simple yet efficient way to improve the robustness of speech recognition. One common use is to create an initial model for bootstrapping. It is accomplished by pooling data from different acoustic environments, similar to the scheme of pooling data from different speakers to train speaker-independent systems.

There are two issues that need special handling in applying multi-style training. First, multi-style training is effective in boosting noise robustness for the cross-condition experiments [13] but usually at the expense of certain performance degradation for the matched conditions. Careful realization ought to be applied to reduce this negative impact. Second, there is a lack of knowledge on the optimal numbers of environments. Without prior knowledge about the test condition, it is not clear how many different acoustic conditions are sufficient to achieve environment independence.

For speech-to-speech applications running on a portable device, babble speech is appropriate as ambient noise for acoustical degradation. In this paper, we carry out our experiments in the context of environmental degradation due to various speech babbles. The acoustical degradation is characterized in terms of signal noise ratio (SNR). Furthermore, we deliberately choose different SNRs for both training and test data to preserve acoustical mismatch between training and test conditions.

3.2. Codebook Dependent Cepstral Normalization

Unknown additive noise and unknown linear filtering (convolutional noise) are two important sources for acoustical degradation. Figure 2 depicts an environment model for acoustical degradation with these two noise sources.

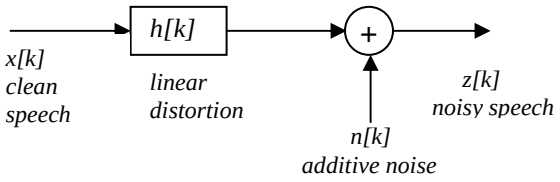

Figure 2: A model of environmental distortion for additive noise and linear filtering

The model can be expressed as equation (1) in terms of cepstral vectors $\boldsymbol{x}$, $\boldsymbol{n}$, $\boldsymbol{z}$, and $\boldsymbol{q}$, where these are for clean signal, additive noise, noisy observed signal, and linear distortion respectively.

$$\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{q} + \boldsymbol{r}(\boldsymbol{x}, \boldsymbol{n}, \boldsymbol{q}) \qquad (1)$$

where the additive vector

$$\boldsymbol{r}(\boldsymbol{x}, \boldsymbol{n}, \boldsymbol{q}) = IDFT[\ln(1 + e^{DFT[\boldsymbol{n}-\boldsymbol{q}-\boldsymbol{x}]}] \qquad (2)$$

represents the joint effects of additive noise and linear filtering.

The CDCN [7,9] algorithm attempts to reverse the effects of linear filter in a cepstral vector of q and the additive noise in cepstral vector, $\boldsymbol{n}$. It first estimates the parameters, $\boldsymbol{q}$, and $\boldsymbol{r}(\boldsymbol{x},\boldsymbol{n},\boldsymbol{q})$ using ML parameter estimation. Then it estimates the uncorrupted cepstral vector, $\boldsymbol{x}$, given the corrupted observation vector, $\boldsymbol{z}$, and the previously calculated $\boldsymbol{q}$ and $\boldsymbol{r}(\boldsymbol{x},\boldsymbol{n},\boldsymbol{q})$ using a MMSE criterion. Equation (3) denotes the estimated clean data given the noisy observation $\boldsymbol{z}$, the linear filter parameter $\boldsymbol{q}$, the correction vectors $\boldsymbol{r}$, and mixture weights $f[l]$,

$$\hat{\boldsymbol{x}} = \sum_{l=0}^{L-1} f_i[l] * (\boldsymbol{z} - \hat{\boldsymbol{q}} - \hat{\boldsymbol{r}}[l]) \qquad i=0,1,\ldots,N-1 \qquad (3)$$

4. EXPERIMENTS AND RESULTS

Experiment Setup. The experiments carried out to evaluate the environmental robustness of the IBM MASTOR system [1] are conducted in a DARPA force protection domain. The source language for translation is English and the target language is Mandarin Chinese. The sentences from the DARPA force protection domain are relatively conversational, interactive, and less-constrained in sentence syntax compared with dictation applications. The language model is created by interpolating a general-domain LM with in-domain LM.

One important goal of this paper is to investigate the correlation of acoustic robustness with translation performance in speech-

to-speech translation applications. The improvements on the acoustic robustness are measured in the context of speech recognition as well as speech-to-speech translation.

Performance Measure

For speech recognition, we use word error rate (WER), a commonly used measure for languages like English. However, it is more complicated to evaluate the performance of translation part. Intuitively, the translation quality is to be determined by the fluency and adequacy of the translation output. This needs multiple human judges and the scores are subjective. When any changes are added into the system, different translation results require a similarly huge amount of labors even for the same test data.

Recently, IBM proposed an objective measure, BLEU [2], to evaluate the translation performance for simplicity and fast turnaround time. As shown in Equation (4), BLEU measures the translation quality based on N-gram probabilities and brevity between hypothesis and reference sentences. The BLEU score is in the range of 0 and 1, where 1 represents a perfect matched translation and 0 means an entirely mismatched translation.

$$\text{BLEU} = \exp(\sum_{\text{i}} \sum_{\text{n}} w_n * \log(P(w_i \mid w_{i-1},.., w_{i-n+1}))) * BP \quad (4)$$

where $P(w_i|w_{i-1},\ldots,w_{i-n+1})$ is the n-gram probability, w_n is the weight, and *BP* is the brevity penalty.

For speech-to-speech translation, we evaluate the translation output using the BLEU score. The performance of speech translation is evaluated using the recognized transcription as input text. For comparison, we also yield the translation result using the correct transcription as input text.

4.1. Multi-Style Training

Noise is digitally added to the training data using recordings from ten different environments to simulate different acoustic environments. These recordings consist speech babble from cafeteria, classrooms, conference halls with lots of background speech. Each recording is 5 minutes long. We splice these ten recordings into one single file and noise samples are randomly selected from it during mixing.

When creating noisy data, we choose not to add noise to the clean data at a fixed SNR. Instead, they are created by adding noise at a fixed magnitude scale. As the speech power varies from sentence to sentence, the resultant data have a wider range of SNRs. For categorization, we do specify the global SNR of the noisy data. Two global SNRs are used, 20dB and 10dB, for multi-style training. The clean training data set has 352,000 sentences from 3200 speakers. A subset of the training corpus consisting of 160,000 sentences from 1560 speakers is used to generate noisy training data. The noisy data are added only in estimating mixture Gaussian distributions while the phonetic decision trees are calculated using the clean data.

A set of test data was collected under a clean environment for the DARPA force protection domain. The test corpus has 1500 sentences from 10 speakers. We deliberately select different global SNRs from those used for training data to generate noisy test data. One is 15dB and the other 8dB.

System	Test Data Condition		
	clean	15dB	8dB
BASLINE	7.23	18.9	44.7
MST-20dB	7.07	15.5	34.8
MST-10dB	7.42	15.4	31.6

Table 1: Speech recognition error rates for baseline and two multi-style training

Table 1 lists the speech recognition results for multi-style training using on all 1500 test sentences. The systems obtained using multi-style training at global SNRs of 20dB and 10dB are denoted as MTS-20dB and MTS-10dB, respectively.

First, the baseline system degrades severely in adverse environments where speech babble is added. Not surprisingly, the recognition error rate decreases for noisy test data when the technique of multi-style training is employed. The MST-10dB system yields a WER of 15.4% (a 19.5% error reduction) for noisy test data of 15dB while a minor performance degradation is observed for clean test data. The MST-20dB system produces a WER of 15.5% for the 15dB noisy data but no performance degradation is observed for the clean data..

System	Test Data Condition		
	clean	15dB	8dB
Correct Text Input	0.630		
BASLINE	0.545	0.430	0.275
MST-20dB	0.551	0.452	0.319
MST-10dB	0.553	0.457	0.329

Table 2: Speech to speech translation performance in terms of BLEU scores for baseline and two multi-style training

Table 2 compares with speech-to-speech translation performance between different systems. The BLEU scores are computed using 4 different references [2]. A BLEU score of 0.63 is produced when the correct text is used as input to translation module. It also shows that the improvement of speech recognition usually translates into higher BLEU scores.

4.2. CDCN

We apply CDCN to the front-end as a pre-processing component in all systems in Section 4.1. Table 3 lists the speech recognition results when CDCN is applied to the test data. It is shown that CDCN is effective in compensating for the noise for both 15dB and 8dB test corpora. The most interesting part is that CDCN also improves the recognition accuracy for clean test data by 7% relatively when used in baseline system.

System	Test Data Condition		
	clean	15dB	8dB
CDCN + BASLINE	6.72	13.6	28.4
CDCN + MST-20dB	6.72	13.2	26.0
CDCN + MST-10dB	7.68	13.9	25.6

Table 3: Comparison of CDCN results in speech recognition

When applied to the two multi-style training systems, CDCN yields the best results. A 25% and a 19% relative error reduction are observed for the 8dB test data for MST-20dB and MST-10dB systems, respectively. In fact, without the multi-style training, CDCN itself provides a substantial performance

improvement of 36% relatively. Similarly, the improvement of noise robustness is also observed in speech-to-speech translation shown in Table 4.

System	Test Data Condition		
	clean	15dB	8dB
CDCN + BASLINE	0.545	0.474	0.364
CDCN + MST-20dB	0.552	0.478	0.380
CDCN + MST-10dB	0.549	0.479	0.378

Table 4: CDCN performances in speech-to-speech translation

4.3. Noise Immunity of WER and BLEU

Figure 3 and Figure 4 compare the best system, "CDCN+MST-20dB", with baseline in the context of both speech recognition and speech-to-speech translation, respectively. As expected, WER increases and BLEU drops when the speech babble noise intensifies. It also reveals the improvements resulting from multi-style training and CDCN for speech recognition are mostly carried over to the speech translation performance.

The trend of the change in BLEU is more of interest than the absolute magnitude of change of BLEU. For example, "CDCN+MST-20dB" outperforms baseline for 8dB test data with a WER of 26.0% versus 44.7% in speech recognition. The corresponding BLEUs are 0.380 versus 0.275. In other words, a 18.7% absolute error reduction in WER is translated into an increase of 0.105 in BLUE. Likewise, when the test data changes from "clean" to "15dB", the system "CDCN+MST-20dB" exhibits an absolute 6.48% change in WER and a change of 0.074 in BLEU.

It is interesting to note that improvement on WER does not guarantee improvement on BLEU when the change is relatively small. For example, CDCN improves the WER of baseline system for clean data from 7.23% to 6.72% but no improvement in BLEU is observed. A careful examination of components in computing BLEU, N-gram coverage does increase, as expected, but the brevity penalty increases. This reflects the fact that when CDCN is used, less insertion errors are generated during recognition. Nevertheless, the general observation is that the more robust the speech recognition is, the better the overall performance can be achieved.

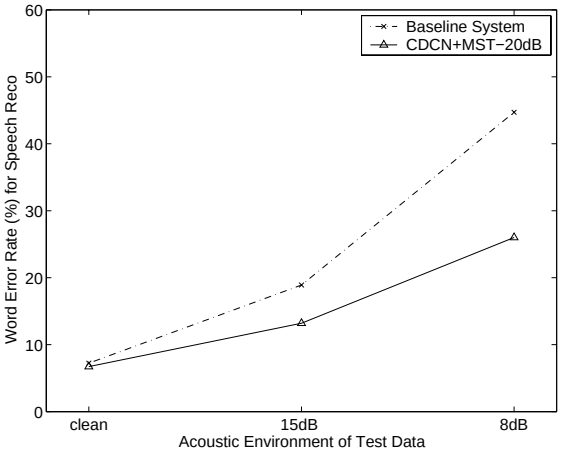

Figure 3: Comparison of word error rates (WER) for speech recognition between baseline and the best system

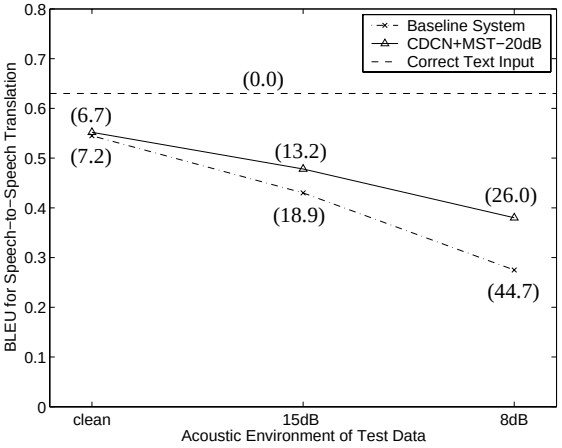

Figure 4: Comparison of BLEU scores between the same systems in Figure 3; numbers in parentheses indicate the corresponding WER in %

5. SUMMARY

Severe performance degradation is observed due to acoustical distortions of babble noise when a speech-to-speech system is used for English-to-Chinese translation. We evaluate the degradation at both speech recognition and speech-to-speech translation scopes. Two techniques, multi-style training and CDCN, are shown effective in compensating for environmental degradations for a speech-to-speech translation application. The best result is achieved by combining CDCN with MST-20dB, which has a 6.72% and a 13.2% of WER for the clean and 15dB test data, respectively. We also study the relation of WER in speech recognition and BLEU in speech-to-speech translation performance.

6. REFERENCES

[1] F.-H. Liu, et al, "Use of Statistical N-Gram Models in Natural Language Generation for Machine Translation", *ICASSP-2003*, 2003.

[2] K. Papineni, et al, "Bleu: A Method for Automatic Evaluation of Machine Translation", *Research Report RC22176*, IBM, Sept. 2001.

[3] A. Berger, et al, "A Maximum Entropy Approach to Natural Language Processing", *Computer Linguistics*, Vol. 22, No. 1, pp. 39-71, 1996.

[4] A. Lavie, et al, "Janus-III: Speech-to-Speech Translation in Multiple Languages", *Proceedings of ICASSP-97*, 1997

[5] W. Wahlster, ed., *Verbmobile: Foundation of Speech-to-Speech Translation*, Springer, 2000.

[6] B. Zhou, et al, "Statistical Natural Language Generation for Speech-to-Speech Machine Translation Systems", *ICSLP-2002*, 2002.

[7] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, Boston, MA, 1993.

[8] R. Donovan, et al, "Phrase Splicing and Variable Substitution Using the IBM Trainable Speech Synthesis System", *ICASSP-1999*, pp. 373-376, 1999.

[9] F.-H. Liu, et al, "Efficient Joint Compensation of Speech for the Effects of Additive Noise and Linear Filtering", *ICASSP-92*, pp. 865-868, 1992.

[10] H. Ney, et al, "Algorithms for Statistical Translation of Spoken Language", *IEEE Trans. on Speech and Audio Processing*, vol.8, no.1 , January 2002

[11] T. Takezawa, et al, "A Japanese-to-English Speech Translation System: ART-MATRIX", *ICSLP-1998*, pp. 2779-2782, 1998.

[12] K. Davies, et al, "The IBM Conversational Telephony System for Financial Applications", *EuroSpeech-1999*, pp.275-278, 1999.

[13] R. Lippmann, et al, "Multi-Style Training for Robust Isolated-Word Recognition", *ICASSP-87*, pp. 705-708, 1987