Improvements in Transcribing Lectures and Seminars

L. Lamel, H. Schwenk, J.L. Gauvain, G. Adda, E. Bilinski

LIMSI-CNRS, BP 133
F-91403 Orsay Cedex, France

Introduction

This paper describes recent research carried out in the context of the FP6 Integrated Project CHIL (chil.server.de) on developing a system to automatically transcribe lectures and seminars. Widely available corpora were used to train both the acoustic and language models, since only a small amount of CHIL data was available for system development. For language model training, text materials come from a variety of online conference proceedings and a neural network language model has been used to take better advantage of the limited data.

Recognizer Overview

The speech recognizer uses the same core technology as the LIMSI Broadcast News Transcription system. Since no CHIL specific audio training data were available, the following corpora were used to train the acoustic models: TED speeches, ISL, NIST and ICSI meetings (total of 96h). The baseline back-off language models were trained on the transcriptions of the audio data (1.1M words) and 18972 articles from speech-related workshops and conferences (35.4M words). The use of transcriptions of conversational telephone speech or even Broadcast News was not very useful. The word-list has 35k words and the OOV rate is 0.17% on the CHIL Jan'05 test set. Word recognition is performed in two passes: initial hypothesis generation followed by an adapted decode. Each decoding pass generates a word lattice which is expanded with a 4-gram language model (LM) or a neural network LM for the last decoding pass.

Experiments and Results

Experimental results are reported on the data used for the Jan 2005 CHIL technology benchmark set, comprised of five seminars (5 speakers, with German, American, Italian and Indian accents). Two of the five seminars were split into development and test portions, and the remaining three were only used for testing purposes. In total there was about 0.75h of development data and 2.1h of test. All data were manually transcribed and segmented. Speech recognition tests were carried out on both the close-talking microphone data and far-field microphone data. For the far-field task, the data from the individual microphone channels could be used, as well as the result of a delay-and-sum beam-forming performed at the University of Karlsruhe.

As shown in Table 1, the word error rate on the CHIL data using the LIMSI RT04 BN system was over 40%. Using more appropriate acoustic and language models, the word error rate is decreased to 26%. Adding just a small amount

	BN models		Without Adapt Data		With Adapt Data		Farfield
	Perplexity	Werr	Perplexity	Werr	Perplexity	Werr	Werr
Back-off LM	275	42.2%	122.1	26.0%	107.5	23.6%	51.9%
Neural LM	-	-	110.2	**24.4%**	99.5	**22.6%**	-

Table 1. Result summary for the unmodified BN system, the CHIL system (with and without using adaptation data) and the CHIL system for farfield microphones.

of speech (1 hour total) from some of the test speakers to the almost 100 hours of audio data results in a word error rate of 23.6%. The word error rate on the beam-formed microphone array data is about twice that obtained on the close-talking microphone data.

The use of a connectioninst LM [2], shown to be performant when LM traning data is limited, has been recently explored for this task. The basic idea is to project the word indices onto a continuous space and to use a probability estimator operating on this space. Both task are performed by a neural network. This is still a n-gram approach, but the n-gram LM probabilities are "interpolated" for any possible context of length n-1 instead of backing-off to shorter contexts. Since the resulting probability densities are continuous functions of the word representation, better generalization to unknown n-grams can be expected. Due to the small size of the adaptation corpus (13.5k words) the neural network LM was trained only on the transcriptions of the audio data and the proceedings so as to ensure better generalization. Results with the neural network LM interpolated with the corresponding back-off LM are given in Table 1. The neural network LM achieved significant improvements with respect to the reference back-off LM: a word error reduction of 1.6% absolute when no biased development data is used, and 1.0% absolute when the development data was used for the back-off LM only. Rescoring the lattices is done in less than 0.3xRT.

Conclusions

The general task of transcribing lectures and seminars is a challenging one, combining the difficulties encountered in processing spontaneous speech with the difficulties of far-field speech recognition. Most of the results reported here are for close-talking microphone data since there was no far-field microphone data available for system development. It is our belief that most of techniques which improve this condition will also improve far-field speech recognition (the original BN system had a word error rate of over 70% on the far-field data). The described recognizers achieve word error rates around 23% without the use of task specific audio or textual training data. A neural network LM was also employed, improving the word error rate by up to 1.6% absolute. Ongoing work is addressing the problem of automatic partitioning of the data into speaker turns and using multi-microphone training data to improve the far-field recognition.

References

1. J.L. Gauvain, L. Lamel, G. Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication*, **37**(1-2):89-108, May 2002.
2. H. Schwenk. Efficient training of large neural networks for language modeling. In *IJCNN* , pp. 3059–3062, 2004.