OPTIMIZING ACOUSTIC MODELS FOR COMMERCIAL SPEECH RECOGNITION USING FOREGROUND SCORES AND DATA WEIGHTING

Daniel Boies, Brian Strope, Mitchel Weintraub, Su-Lin Wu
{boies, bps, mw, swu}@nuance.com

Nuance Communications, 1005 Hamilton Ct., Menlo Park, CA, 94025

ABSTRACT

This paper describes a data-driven technique for optimizing the acoustic models for speech recognition systems that target commercial applications over telephones. Frame-averaged foreground log-likelihoods (foreground scores) correlate to recognition errors. These scores are used together with gender to optimize data weighting for the acoustic model. This process is interpreted as increasing the priors and associated parameters for poorly modeled data. The score-based optimization leads to about 7% fewer semantic errors on a live evaluation set collected after the last data used to estimate the acoustic model.

1. INTRODUCTION

Acoustic models for commercial speech recognition integrate large amounts of information. In the last 5 years, CPU speed and memory sizes have increased to the point where detailed acoustic models for real-time large-vocabulary systems can include parameters targeted for specific challenges, often without significantly degrading recognition accuracy on median data. General-purpose models can be built which include optimizations for specific acoustic conditions, for example, hands-free data in the car, non-native talkers, different cellular codecs and microphone types. Models can also include optimizations for varying application types, which impose different language modeling requirements, for example, digit strings, stock quotes, name lists, or call-routing tasks. As these lists grow, balancing for the different acoustic and language modeling combinations to build a general-purpose acoustic model implies an iterative, manual process that does not scale well as the available training data and supportable model sizes increase. Here we propose using a simple per-utterance statistic that correlates to recognition error rates to help automate the optimization.

Approaches for optimizing acoustic models have argued that flaws in modeling assumptions motivate tying the optimization more closely to recognition error rates. Many of the techniques are based on replacing maximum likelihood estimation (MLE) with maximum mutual information estimation (MMIE) [1]. This approach uses both the MLE constraint of maximizing the likelihood of the observation given the transcribed word sequence and a simultaneous constraint of minimizing the log-likelihood of the observation given all possible word sequences. Efficient MMIE implementations have been developed for large vocabulary continuous speech recognition (LVCSR) [2,3]. Other approaches have used minimum classification error (MCE) together with simultaneous feature and model optimizations [4]. Usually discriminatively trained systems have shown the best improvements on the tasks considered during training. However training optimizations, including relaxing the language model constraints on the list of all possible word sequences and adjusting the relative influence of the language model and the acoustic model, have led to improvements that generalize across LVCSR tasks [5,6].

Other improvements over MLE training have been made based on boosting theory, which combines sequences of classifiers where each successive classifier improves on the errors of the predecessor [7]. Techniques have been developed to apply boosting to LVCSR [8], and consistent with the current work, other efforts have used these ideas with utterance-specific measures [9].

Considering the wide range of language modeling requirements (from trivial to near-conversational) and the variety of acoustic challenges for current commercial speech recognition over telephones, we implemented a conservative simplification of other discriminative training techniques. As with other approaches, model parameters are concentrated for challenging acoustic conditions in a data-driven manner. But instead of imposing even a relaxed language model and directly using incorrect matches, we use an acoustic error correlate to identify problem utterances. Finally instead of a frame-based optimization we simply increase the weight of potential problem utterances when estimating model parameters.

Section 2 describes the foreground score statistic. Section 3 describes how this statistic is used to help optimize the models. Section 4 describes the experiment results.

2. FOREGROUND SCORES

Incorrectly assuming independence of the observation sequence, most speech recognition systems accumulate the sum of the log-likelihoods for observing each frame given a specific recognition hypothesis. During testing, the hypothesis with the highest accumulated log-likelihood, or score, is the recognized result. During maximum-likelihood training, the hypothesis is the transcription. The accumulating (forward and backward) scores are used to estimate the probability of observing specific frames in the underlying model states. The training process maximizes the scores given the training data.

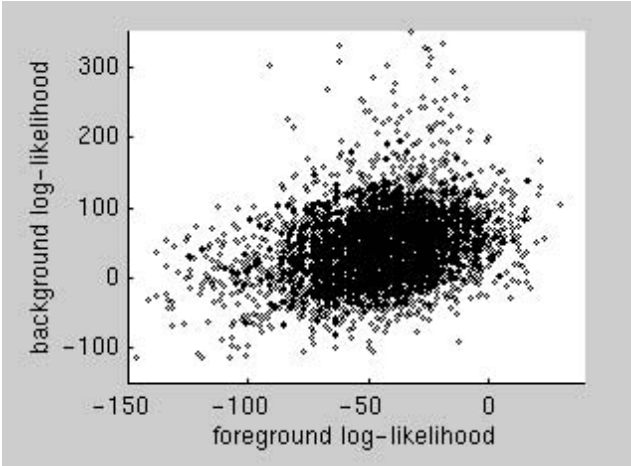

Fig. 1. Scatter-plot of foreground and background scores.

MLE training solutions for state alignments and model clustering techniques for optimizing and sharing GMM parameters [10] achieve local maxima by increasing the scores for more common data. To do that, model parameters are concentrated for common data. Less common data has lower accumulated scores, and higher error rates. If there are robust error improvements to be made by intentionally moving the models away from the MLE solution, then lower scores might help identify the problem utterances.

Instead of considering the complete accumulated score from a forced alignment, we separate frames that are aligned to background models (usually non-speech) from those aligned to foreground models (the transcription). The scores for the foreground and background frames are averaged across the utterance to provide two measurements per utterance: the foreground score and the background score.

Figure 1 shows a scatter plot of foreground and background scores for a sub-sample of the training set measured using forced-alignments with a gender-balanced baseline acoustic model. There is only a weak correlation (0.2) between the scores. Intuitively, some utterances have background segments that are not well modeled, and others have challenging foregrounds.

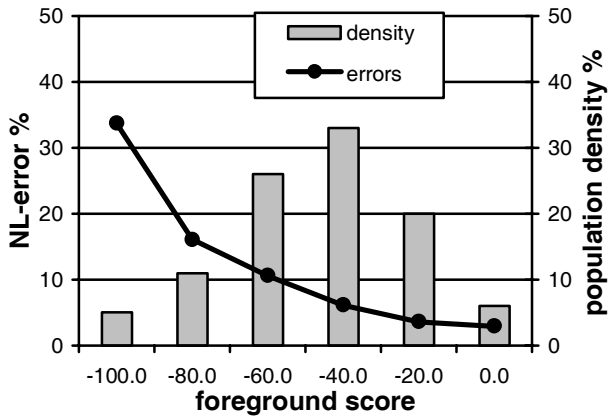

Fig. 2. Error rates as a function of foreground scores.

Figure 2 shows recognition error rates using a previous acoustic model on a task where the data is separated into bins by foreground scores. More details about the tests and the error measures are given in Section 4. The general trend is that foreground scores provide some prediction of recognition errors. Utterances with scores in the lowest 5% lead to just over 10 times the number of errors of the highest scoring 6%.

3. OPTIMIZING THE MODELS

A simple data-weighting strategy is used to optimize the acoustic models. Other more elaborate data-weighting techniques were explored (including jack-knifing the training data), but none out-performed the direct method described here.

First, a baseline acoustic model is trained from boot models using gender-corrected data weighting so that each gender has equal representation in the model. Second, the baseline model is used to force-align the training data. Third, foreground scores are computed for each utterance and the training data is separated into the lowest scoring 20% and the highest scoring 80%. Finally, a single training iteration is used to re-estimate the model parameters using a 10x increase on the data weighting for the worst 20% data. This provides the final weighted model.

Increasing the data weight for the lowest-scoring data increases the models' representation of the prior for that type of data. Similarly by increasing the relative variance of the training counts associated with that data, the weighting increases the number of model parameters associated with the worst 20% of the data.

4. EXPERIMENT RESULTS

4.1. Semantic error measure

All experiments here measure only utterance-level natural-language understanding errors, or NL-errors, not word errors. While these are strongly correlated, NL-errors de-emphasize the significance of word errors on filler words that would not change the progression of the application.

For example, consider the transcription "I'd like one two three" for a dialog state where the caller is prompted for a flight number. The recognition result "flight one two three" would not lead to an NL error, but "I'd like one two eight" would.

4.2. Testing data and experiment setup

The experiments reported here do not include recognition rejection, and only use utterances that are parsed by the test grammar. All tests are with American English.

Three probabilistic finite state grammars were used. The grammars were pieced together from fielded deployments. The first grammar included stock quotes and common main-menu items like "help." The second was a flat digit grammar that allowed up to 12 digits and included common surrounding filler phrases. The third was a general confirmation grammar with common yes/no variations, again including filler phrases.

Before retraining the acoustic models, we mined recent data to build test sets. Our existing acoustic models were used to get foreground scores for transcribed data that had been collected since the models were released. From the score distribution we chose score thresholds that led to 6 bins (fg0 to fg5) ranging from the lowest 5% to the next 11%, 26%, 33%, 20%, and the highest 6%. For each bin of data we found a similar number of utterances that could be parsed by each of the three grammars. The total test set included about 25k utterances. This was a development test.

The graphs below average errors across the three grammars and show the performance for each bin (fg0 to fg5). The total error measure weights the errors across the bins by the percentages above that reflect the population score density.

The development test was used to fit two parameters: the data separation threshold (20%) and the increased weighting (factor of 10). It also seems reasonable to question whether problems from less common utterances are stable with time. Therefore, an evaluation test was built. After the new acoustic models were finished, we mined data, collected since we built the training lists for the new models, to build a similar score-based test set. Unlike the development set, the evaluation set was collected after the training data and was typically from a different set of deployments.

4.3. Recognizer

All experiments below used the commercially available Nuance 8 recognizer together with the prototype acoustic models described here.

4.4. Results

Figure 3 shows NL-error rates for the baseline acoustic models and the weighted acoustic models on the development set. The baseline models were gender-balanced using data weighting, and the "weighted" models added the score-based process described in Section 3 together with the gender balancing.

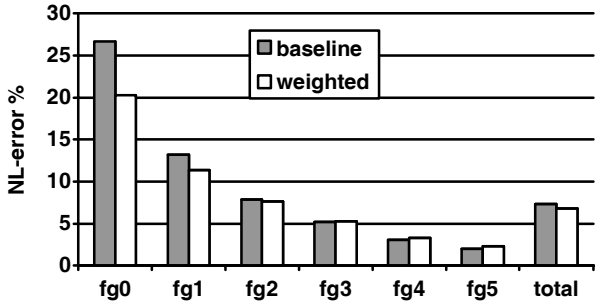

Fig. 3 Dev-set NL-error: baseline, and weighted**.**

Figure 4 compares the same two models on the evaluation test-set. As with the development set, fg0-fg2 are improved with small degradations in the other tests. Overall both tests show the same 7% relative reduction of NL-errors.

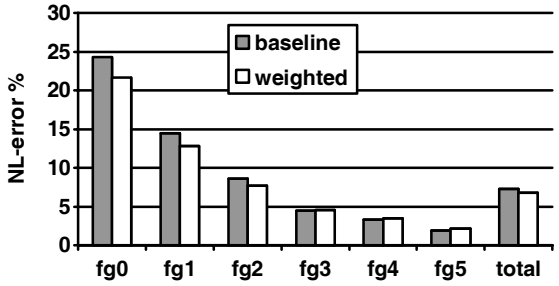

Fig. 4 Eval-set NL-error: baseline, and weighted.

Figure 5 compares the 10x-weighted model to two models that move further from the baseline model. With the first model, all parameters are re-estimated from a single iteration using only the worst 20% data. Compared to the 10x-weighted model, this is the same as infinite weighting on the worst data. With the second model, we retrained from boot models and relearned model structure (allophonic clustering and Gaussian allocation) using only the worst 20% of the data.

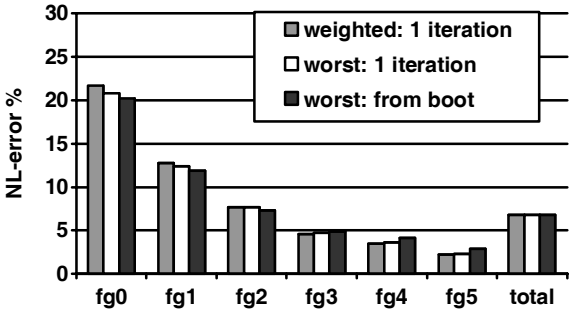

Fig. 5 Eval-set NL-error: 10x weighted, worst 20% alone, and worst 20% from boot models.

The total performance is the same with any of these. In general, a few errors are moved from the low scoring data to the higher scoring data. If there is a large application cost associated with high recognition errors rates in pathological conditions, then "flattening" the error distribution without reducing errors might also be an interesting trade-off.

5. DISCUSSION

This paper describes using frame-averaged foreground log-likelihoods (foreground scores) to optimize acoustic models for commercial speech recognition on telephones. Foreground scores based on forced alignments are used to identify low-scoring utterances. The relative weights for the lowest 20% are increased by a factor of 10 for a final training iteration. This technique leads to 7% fewer errors on a live evaluation test collected after the data used to train the models.

Using an error-correlate to drive modeling optimizations is a compromise between hand-tweaked data weighting of difficult labeled utterance types and more direct discriminative training. It targets unspecified modeling challenges in a data-driven manner, while using all available training data, and without imposing test grammars for training optimizations.

6. REFERENCES

[1] L. Bahl, P. Brown, P. de Souza, and R. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," *Proc. ICASSP '86*, Tokyo, pp. 49-52, 1986.

[2] L. Bahl, M. Padmanabhan, D. Nahamoo, and P. Gopalakrishnan, "Discriminative training of Gaussian mixture models for large vocabulary speech recognition systems," *Proc. ICASSP '96*, Atlanta, vol. 2, pp. 613-616, 1996.

[3] V. Valtchev, J. Odell, P. Woodland, and S. Young, "MMIE training of large vocabulary speech recognition systems," *Speech Communication*, vol. 22, pp. 303-314, 1997.

[4] M. Rahim, Y. Bengio, and Y. LeCun, "Discriminative feature and model design for automatic speech recognition," *Proc. Eurospeech '97*, Athens, vol. 1, pp. 75-78, 1997.

[5] D. Povey and P. Woodland, "Improved discriminative training techniques for large vocabulary continuous speech recognition," *Proc. ICASSP '01*, Salt Lake City, 2001.

[6] R. Cordoba, P. Woodland, and M. Gales, "Improved cross-task recognition using MMIE training," *Proc. ICASSP '02*, Orlando, pp. I-85-88, 2002.

[7] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119-137, 1997.

[8] G. Zweig, and M. Padmanabhan, "Boosting Gaussian mixtures in an LVCSR system," *Proc. ICASSP '00*, Istanbul, pp. 1527-1530, 2000.

[9] C. Meyer, "Utterance-level boosting of HMM speech recognizers," *Proc. ICASSP '02*, Orlando, pp. I-109-112, 2002.

[10] V. Digalakis, P. Monaco, and H. Murveit, "Genones, generalized mixture tying in continuous hidden Markov model-based speech recognizers," IEEE SAP, vol. 4, no. 4, pp. 281-289, 1996.