Confidence Scoring and Rejection using Multi-Pass Speech Recognition

Vincent Vanhoucke

Nuance Communications, Menlo Park, CA, USA
vincent@nuance.com

Abstract

This paper presents a computationally efficient method for using multiple speech recognizers in a multi-pass framework to improve the rejection performance of an automatic speech recognition system. A set of criteria is proposed, which determine at run time when rescoring using a second pass is expected to improve the rejection performance. The second pass result is used along with a set of features derived from the first pass to compute a combined confidence score. The feature combination is optimized globally based on training data. The combined system significantly outperforms a simple two-pass system at little more computational cost than comparable one-pass and two-pass systems.

1. Introduction

The determination of whether a recognition hypothesis H is the correct one based on the input speech signal O can be performed with reasonable accuracy by computing an estimate of its posterior probability:

$$p(H|O) = \frac{p(O|H)p(H)}{p(O)}$$

For computational reasons, this posterior is typically an approximation derived using statistics collected during the decoding of the speech [3].

Two different avenues have shown to be very promising in terms of improving the confidence scores derived from the acoustic posterior:

1. The combination of the scores with other statistics derived from the recognizer [6],
2. The combination of the outputs of multiple recognizers [4, 5].

The combination of multiple recognizers is a very powerful technique, but can be very expensive computationally if all the recognizers are to be run in parallel for any given utterance. It is more efficient to cascade the systems into a multi-pass framework, where an inexpensive recognizer is run as a first pass, followed by a rescoring of its output using a set of more detailed second passes. By limiting the number of utterances actually exercising this rescoring step, the average computational expense of the overall system can be constrained to be just slightly above the cost of a one-pass system, with a much improved overall accuracy [7].

When it comes to improving the rejection performance of the system, however, the optimization of a multi-pass system has to obey a different set of requirements than those of a system optimized for accuracy alone. For example, a system optimized for accuracy will want to determine which recognizer (the first pass or second pass) has produced a correct answer. If the first pass is expected to be correct, then the second pass doesn't have to be run. When it comes to confidence, however, the output of both recognizers is potentially of interest, since a disagreement between the two outputs could indicate that the result is questionable. The fact that the system determined that it needed to run a second pass is also an indication that the confidence in the result of the first pass should be low.

In the following, we will show how the logic governing a multi-pass system optimized for accuracy can be enhanced to take into account the requirements of improving rejection. Section 2 describes a simplified baseline multi-pass system. Section 3 introduces a set of multi-pass features which can be added to the computation of the confidence score. Section 4 introduces a modification to the multi-pass logic which further improves the rejection performance.

2. Baseline Multi-Pass System

The baseline system we consider uses a very simple multi-pass strategy, depicted in Figure 1.

A first pass recognizer (Pass I) is run, and a set of hypotheses is produced, along with an acoustic posterior P for each of them. An additional likelihood margin D is computed, which measures how close the alternative hypotheses are from the best hypothesis. A small D indicates that there are possible alternatives to the best hypothesis in the grammar, and hence that a rescoring with better models could help disambiguate between them. A large D indicates that there is a lack of likely competing hypotheses in the grammar, which makes any attempt at rescoring to improve the recognition result superfluous. Based on a threshold Δ, we determine whether it is necessary or not to run a rescoring pass (Pass II). Finally, a threshold Θ is applied to the posterior P of the best hypothesis to determine whether to accept or reject it. This threshold is determined by the operating requirements of the system in terms of correct accept (CA) rate vs. false accept (FA) rate.

3. Multi-Pass Confidence Features

3.1. Second Pass Confidence Penalty

The two measures of confidence (Acoustic Posterior P and Likelihood Margin D) provide different information about the top recognition hypothesis. P evaluates whether it is a good acoustic match, while D measures whether other hypotheses would be good candidates as well, and thus whether the recognizer picked its top hypothesis by chance.

A simple but effective method for combining the two sources of information is to define a penalty:

$$\delta\Theta = \Theta_2 - \Theta_1$$

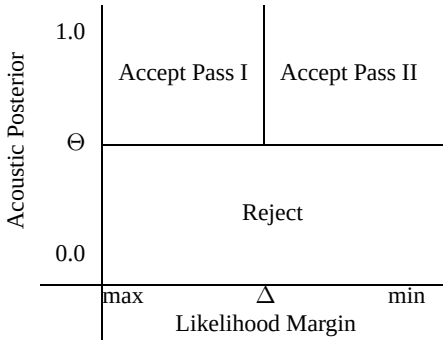

Figure 1: *Baseline system*

which is applied to the confidence score of the utterances which are rescored using the second pass. This translates into the acoustic confidence threshold for utterances going to the second pass to be increased by $\delta\Theta$, as depicted in Figure 2.

The value of $\delta\Theta$ can be learned simply by collecting the information:

- Acoustic Posterior P of best hypothesis so far,
- $D > \Delta$,
- best hypothesis is correct / incorrect,

for some representative data, and applying Linear Discriminant Analysis (LDA, see e.g. [8, Chapter 4]) to the data to best separate the correct utterances from the incorrect ones. In practice, this optimization provides a very robust estimate of $\delta\Theta$ which is very stable across values of Θ_1, and very consistent across testsets.

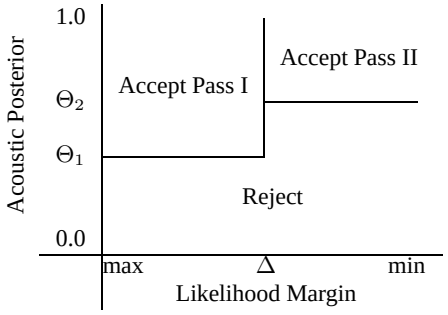

Figure 2: *System with Second Pass Confidence Penalty*

3.2. Recognizer Disagreement Confidence Penalty

When the second pass recognizer is run, it provides an estimate of the recognition result which is relatively independent from the first pass result. As a consequence, whether the two recognizers agree on the resulting output can be expected to be a very salient feature for confidence scoring. As previously, we incorporate this information by calculating another penalty:

$$\delta'\Theta = \Phi - \Theta_2$$

to be applied to the confidence, as depicted in Figure 3.

The value of $\delta'\Theta$ can be learned jointly with $\delta\Theta$ by collecting the information:

- Acoustic Posterior P of best hypothesis so far,
- $D > \Delta$,
- Pass I agrees / disagrees with Pass II,
- best hypothesis is correct / incorrect,

for some representative data, and applying LDA to the data.

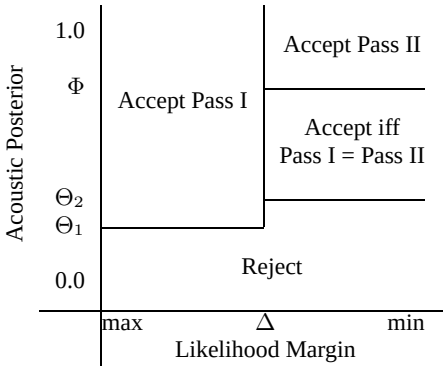

Figure 3: *System with both Confidence Penalties*

4. Multi-Pass Confirmation Logic

Because of the random effect that beam pruning has on utterances which are not well modeled in the search space, it is often observed that an out-of-grammar (OOG) utterance matches a specific recognition hypothesis, possibly with little or nothing to do with the actual speech, much better than others. In this case, the acoustic confidence of the utterance is low, while the likelihood margin is large, instructing the system not to exercise the rescoring pass.

By forcing the second pass to run in such situations, a "second opinion" can be obtained using very distinct acoustic models, which will exhibit a different behavior when presented with OOG data. For simplicity, the same penalty:

$$\delta'\Theta = \Phi_1 - \Theta_1 = \Phi_2 - \Theta_2$$

as the penalty applied in the second pass is applied to the confidence score of those utterances, as depicted in Figure 4. This penalty could as well be learned from data as in the previous cases.

This strategy forces the second pass to run more often than in the baseline, but the additional cost is only incurred when the likelihood of a misrecognition is high.

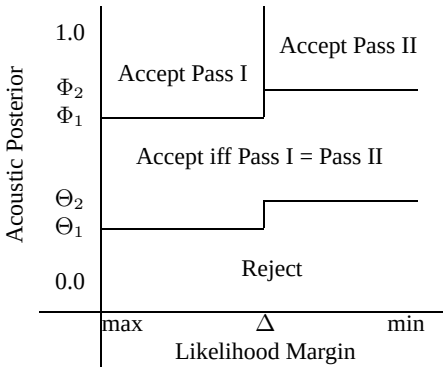

Figure 4: *System with Second Pass Confirmation*

5. Experiments

5.1. Experimental Setup

The experiments were run on a speaker-independent American English system. The first pass recognition engine used is a context-dependent HMM system with 18000 triphones and tied mixtures based on Genones [1]: each state cluster shares a common set of Gaussians called Genone, while the mixture weights are state-dependent. The system uses 2000 Genones and 32 Gaussians per Genone. The models are trained using Maximum Mutual Information Estimation [9], and use Mixtures of Inverse Covariances [2] as a covariance model. The second pass recognizers are one male and one female recognizer with comparable parametrizations, but trained using maximum likelihood. The features are 27 dimensional, including MFCC, Δ and $\Delta\Delta$.

The first test-set is a collection of 50000 utterances from a business listings recognition task, including 20% of OOG data collected on the same task. The second test-set is a collection of 40000 utterances from a variety of tasks, including digits strings, stock quotes, and city names, with 50% of OOG data collected on the same tasks. In each case, the language model is a rule-based grammar specifically built and tuned for the corresponding task.

5.2. Rejection Performance on Business Listings

Figure 5 shows the rejection performance of the various configurations on the business listings testset. The graph depicts the *Correct Accept* (CA) rate , i.e. the number of semantically correct utterances which were accepted by the system as a percentage of the total number of utterances, as a function of the *False Accept* (FA) rate, i.e. the number of semantically incorrect utterances which were accepted by the system, also as a percentage of the total number of utterances. A perfect system would have its CA/FA curve follow the left vertical axis (0% FA) and the top horizontal axis (100% CA).

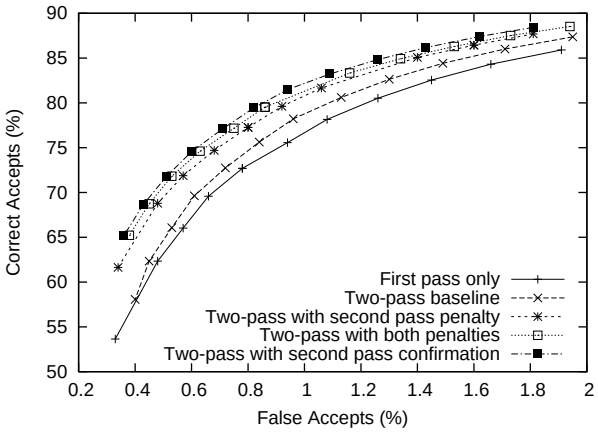

Figure 5: *Rejection performance of the various systems on business listings.*

The best system using second pass confirmation leads to a 3 to 5% absolute improvement in the CA rate at a given FA rate, or conversely a 0.2 to 0.4% absolute reduction in the FA rate at a given CA rate.

5.3. Impact on Decoding Speed

Figure 6 shows the percentage of utterances, represented by the length of the vertical bar, which use the second pass rescoring at each operating point. The system with confirmation will use a second pass rescoring at most 11% of the time on this task, while the baseline system uses the second pass at most 6% of the time. This increase has a negligible effect on the total efficiency of the system.

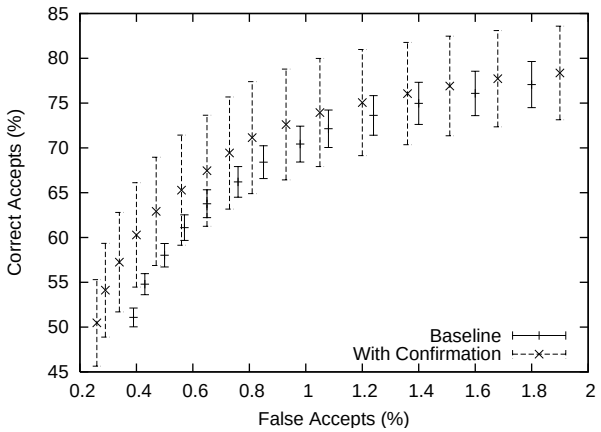

Figure 6: *Percentage of the data going to the second pass (vertical bars) as a function of the operating point.*

5.4. Rejection Performance at High OOG Rate

Figure 7 shows the rejection performance of the various configurations on the mixed testset with 50% of OOG utterances:

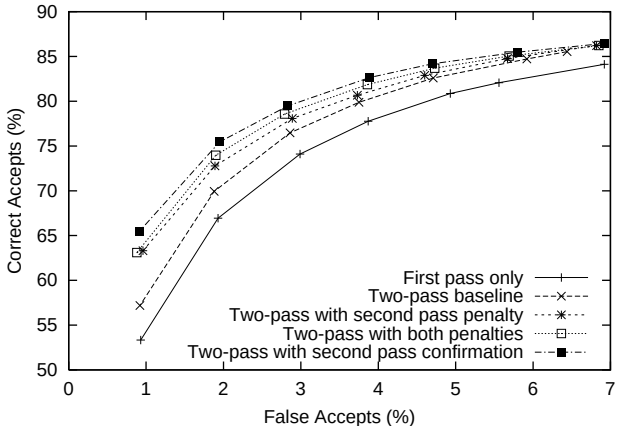

Figure 7: *Rejection performance of the various systems on tasks with high OOG rate.*

At a very low 1% FA rate, the CA rate of the best system is more than 12% (absolute) better than the one-pass system, and more than 8% (absolute) better than the baseline two-pass system. The one-pass system could only achieve the same CA rate at the cost of doubling the number of false accepts.

6. Conclusion

This paper presents a simple, efficient method for leveraging two-pass rescoring for the purpose of improving rejection performance. The system uses a simple decision logic based on the acoustic posterior and likelihood margin of the top hypothesis of the first-pass recognizer to determine whether the rescoring pass should be run. The combined output of the two recognizers determines the final recognition result as well as its confidence score based on a set of penalties, which depend on both the likelihood margin and the semantic agreement between the two passes. The value of these penalties can be derived from data using LDA or other classification methods. The resulting system provides much better rejection performance at a very small computational cost.

7. Acknowledgments

The author would like to thank Brian Strope, Mitch Weintraub and Larry Heck for their input.

8. References

[1] Digalakis, V., Monaco, P. and Murveit, H., "Genones: Generalized mixture tying in continuous hidden Markov model-based speech recognizers," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 4, pp. 281–289, 1996.

[2] Vanhoucke, V. and Sankar, A., "Mixtures of Inverse Covariances," *IEEE Trans. on Speech and Audio Processing*, vol. 12, no. 3, May 2004.

[3] Evermann, G., and Woodland, P.C., "Large Vocabulary Decoding and Confidence Estimation using Word Posterior Probabilities," *Proceedings of ICASSP'00*, pp. 2366-2369, Istanbul, 2000.

[4] Evermann, G., and Woodland, P.C., "Posterior Probability Decoding, Confidence Estimation, and System Combination," *Proceedings NIST Speech Transcription Workshop*, College Park, MD, 2000.

[5] Sankar, A., "Bayesian Model Combination (BAYCOM) for Improved Recognition," *Proceedings of ICASSP'05*, 2005.

[6] Hazen, T.J., Burianek, T., Polifroni, J., and Seneff, S., "Recognition Confidence Scoring for Use in Speech Understanding Systems" *Proceedings of ASR'00*, 2000.

[7] Mao, M.Z., Vanhoucke, V., and Strope, B., "Automatic Training Set Segmentation for Multi-Pass Speech Recognition," *Proceedings of ICASSP'05*, 2005.

[8] Webb, A., *Statistical Pattern Recognition*, Arnold, 1999.

[9] Woodland, P.C., and Povey, D., "Large Scale Discriminative Training for Speech Recognition," *Proceedings of ISCA ITRW Automatic Speech Recognition: Challenges for the Millenium*, pp. 7–16, Paris, 2000.