UTTERANCE VERIFICATION BASED ON STATISTICS OF PHONE-LEVEL CONFIDENCE SCORES

Ananth Sankar and Su-Lin Wu

Nuance Communications
1380 Willow Road
Menlo Park, CA 94025

ABSTRACT

We present new acoustic confidence scores for utterance verification based on novel combinations of phone-level posterior probability statistics. A common utterance acoustic confidence score used in the literature is the arithmetic mean (computed over the utterance) of the phone log posterior probabilities. This approach can be problematic when a large part of the utterance is in-grammar (IG), but a small part is out-of-grammar (OOG). For example, a caller says an OOG name "Larry" and is incorrectly recognized as an IG name "Harry". Since most phones were correctly recognized, the mean of the phone posteriors gives a high utterance level score even though the recognition result should ideally be rejected. We introduce additional statistics, such as the variance and low percentile points of the phone-posterior scores over the utterance, that help in capturing the deviation of otherwise good recognition matches. We report on our experiments on combining these statistics. In particular, by normalizing the mean with the standard deviation, we achieved a 10-20% relative improvement in performance for alpha-digit test sets where OOG utterances are often incorrectly recognized as very similar IG ones.

1. INTRODUCTION

Utterance-level verification is a necessary component for real-world automatic speech recognition (ASR) applications. Acoustically-based methods work by computing an utterance-level confidence score that measures how well the recognition hypothesis matches the observed utterance data. Utterances whose confidence score falls below a pre-determined, application-specific threshold are rejected. Ideally, misrecognized IG utterances and all OOG phrases would be rejected by the utterance verification mechanism.

Various methods have been proposed for computing confidence scores, including purely acoustic measures [1], measures that incorporate language model information (e.g., word graphs and N-best lists [2]), and combined measures [3]. A common utterance-level acoustic confidence score is the geometric average of the posterior probabilities of the phones in the hypothesis, or equivalently, the arithmetic average of the log posterior probability of the phones [4]. In the next section we show how this score relates to the posterior probability of the recognition hypothesis.

While a simple, inexpensive approach, the mean phone posterior probability does not always work well. For example, in some recognition errors, the hypothesis is very similar to what the speaker actually said, but is off by a few phones. Since most of the phones in the recognition hypothesis are correct, and only a few wrong, the mean posterior score will typically be high, resulting in the incorrect hypothesis being accepted. For such cases it makes sense to think of other statistics of the phone posterior probabilities that capture the effect of the outlier phones for which the acoustic match is poor. An example of where this may happen is when a person says an OOG name "Larry", but is misrecognized as "Harry".

In this paper, we propose the use of the variance and low percentile points (for example, the 5th or 10th percentile) of the phone posterior probability scores as new measures of acoustic confidence that can capture the mismatch of a few phones in the recognition hypothesis.

Section 2 motivates and describes our baseline acoustic confidence score, which is the geometric mean of the phone-posterior probabilities over the utterance. In Section 3, we lay out the new statistics that we will study. In Section 4, we consider different ways to combine the statistics to compute utterance-level scores. Section 5 presents experimental results combining various statistics. We show that simply dividing the mean phone posterior score by the standard deviation gives the best results, improving performance by 10-20% in alpha-digit tasks, while not detrimentally affecting other tasks.

2. PHONE-POSTERIOR PROBABILITY-BASED SCORES

The geometric mean of the phone-posterior probabilities [4, 1] can be well motivated as we show below.

It is meaningful to score a recognition hypothesis H with the posterior probability of the hypothesis given the observed feature vector sequence $\boldsymbol{X}$, i.e.,

$$c(H) = P(H|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|H)}{P(\boldsymbol{X})}P(H). \quad (1)$$

Considering only the Viterbi segmentation of the hypothesis, we can write the conditional probability in the right-hand-side of Equation 1 as

$$P(\boldsymbol{X}|H) = P(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_m|h_1, h_2, \ldots, h_m),$$

where $\boldsymbol{X}_i$ is the feature vector sequence segmented into phone h_i. Assuming the phone observations are independent of each other, and depend only on the phone into which they are segmented, we get

$$\begin{aligned} P(\boldsymbol{X}|H) &= \Pi_{i=1}^{m} P(\boldsymbol{X}_i|h_i) \\ P(\boldsymbol{X}) &= \Pi_{i=1}^{m} P(\boldsymbol{X}_i). \end{aligned}$$

Finally, if we use a unigram phone language model for $P(H) = P(h_1, h_2, \ldots, h_m)$, we have

$$\begin{aligned} c(H) &= \Pi_{i=1}^{m} \frac{P(\boldsymbol{X}_i|h_i)}{P(\boldsymbol{X}_i)} P(h_i) \\ &= \Pi_{i=1}^{m} P(h_i|\boldsymbol{X}_i). \end{aligned}$$

Our independence assumption on the m-length phone sequence is quite strong, and in order to correct for it, we may normalize by the number of phones, raising the right-hand-side of the last equation to the power of $\frac{1}{m}$. This gives rise to the method of using the geometric mean of the phone posterior scores:

$$c(H) = (\Pi_{i=1}^{m} P(h_i|\boldsymbol{X}_i))^{\frac{1}{m}}$$

Equivalently, the confidence score in the log probability domain is (we use the same notation $c(H)$ for the log domain too):

$$c(H) = \frac{1}{m} (\Sigma_{i=1}^{m} \log P(h_i|\boldsymbol{X}_i)).$$

Since it is convenient for the score to lie in a fixed range, it is common practice to pass c(H) through a monotonically increasing squashing function, such as a sigmoid. The monotone property makes sure that the rejection behavior is unaltered.

Making similar independence assumptions at the frame level, we compute $P(h_i|\boldsymbol{X}_i)$ as the geometric mean of the frame-level posteriors $P(h_i|\boldsymbol{x}_{i,t})$:

$$P(h_i|\boldsymbol{X}_i) = \left(\Pi_{t \in t_i} \left[\frac{P(\boldsymbol{x}_{i,t}|h_i)}{P(\boldsymbol{x}_{i,t})} \right] P(h_i) \right)^{\frac{1}{|t_i|}},$$

where $\boldsymbol{x}_{i,t}$ is the tth feature vector in the sequence that is segmented into h_i, and $|t_i|$ is the number of frames in the segmentation. In our system, we compute $P(\boldsymbol{x}_{i,t}|h_i)$ using the likelihood score for the corresponding context-dependent phone. For computational reasons, the normalizing term $P(\boldsymbol{x}_{i,t})$ is computed by summing over the context-independent phones [5].

3. UTTERANCE-LEVEL STATISTICS OF PHONE-POSTERIOR SCORES

In the previous section, we laid out the motivation for the use of the arithmetic mean of the log posterior probabilities of the phones in the recognition hypothesis as an utterance-level confidence score. However, as we described in the introduction, when the recognition hypothesis is only slightly different from what the speaker actually said, a large score results, and the system accepts the misrecognized utterance.

To address such cases, we are motivated to look for measures that capture the deviation of a small fraction of phones in the recognition hypothesis. Low percentile points, such as the 5th, 10th, or 20th percentile points are such statistics. They represent the worst scoring phones, but not the high scoring ones. Thus we hope to be able to reject hypotheses that are phonetically almost the same as what was actually said. In a related idea, low scoring words were considered for utterance verification in [6].

Since we sometimes have very few phones in the utterance (as in the case of a Names task), the low percentile estimates will be somewhat noisy. To arrive at a similar score, we also tried using the mean of the scores below a certain percentile point, for example, the mean of the scores below the median.

Another statistic of interest is the variance (or standard deviation). The variance will be larger when there are a few outlier phones with poor acoustic matches. Typically, we would expect an utterance to be rejected to either have a low mean and/or a large variance, whereas an utterance to be accepted will have both a high mean and a low variance. Thus it makes sense to divide the mean by the standard deviation and reject based on this normalized score.

4. COMBINATION OF UTTERANCE-LEVEL STATISTICS

Apart from using the standard-deviation-normalized mean score, we also tried combining the mean with the other statistics by building a simple statistical model on the various scores for the IG and OOG data. A score feature vector $\boldsymbol{S}$ was created with each statistic of interest being a component, and a Gaussian distribution was trained for the IG and OOG utterances. The utterance-level confidence score was then computed using the following log likelihood ratio:

$$c(H) = logN_{ig}(\boldsymbol{S}) - logN_{oog}(\boldsymbol{S}). \quad (2)$$

This is similar to one of the methods suggested in [7] for combining different acoustic scores. We also tried Fischer's linear discriminant analysis [8] to compute the best linear combination of the statistic features.

5. EXPERIMENTAL RESULTS

The speech recognition system employed for our experiments uses Genonic hidden Markov models (HMM). In Genonic HMMs [9], triphone states are clustered using bottom-up agglomerative clustering. Each state cluster shares a set of Gaussians (also called a Genone). Each state in a cluster has an independent set of mixture weights to the Gaussians in the shared Genone. The American English acoustic models used for our experiments contain about 25,000 triphone HMMs, 500 Genones, and 32 Gaussians per Genone, for a total of 16,000 Gaussians.

Our experiments use field data from a variety of real, in-service applications, including variable-length alpha-digits, length-16 digit strings, small-vocabulary menu, medium-vocabulary names, and a large-vocabulary stock quote task. In each grammar, application-specific constraints are applied where possible to optimize the recognition performance. For example, if a certain letter could only appear in a certain position in the alpha-digit string, this constraint was represented in the recognition grammar.

Table 1 shows the approximate number of IG and OOG utterances that were used for each test. The relative sizes do not reflect the actual proportions from a live application. The second column in the table gives the acronym with which we will refer to these test sets.

Table 2 shows the individual utterance-level statistics of the phone-posterior log probability estimates that we experimented with, and the acronyms used in the plot legends.

In Figure 1, we use the ALPHA test set to plot ROC curves for the Mean, NormMean, Pct5, and Mean50 statistics. On the x-axis are the false-accept OOGs (FA_{OOG}), and on the y-axis is the in-grammar error (Error IG), which is the sum of the false-reject IGs (FR_{IG}) and false-accept IGs (FA_{IG}), the latter being the misrecognitions of IG utterances. Figure 2, plots the same curves for

Data Type	Acronym	Number IG	Number OOG
16-digit strings	DIG16	5000	5000
variable-length alpha-digits	ALPHA	1500	2500
menu	MENU	2500	500
names	NAMES	2000	4000
stock quotes	QUOTES	2500	500

Table 1. Number of utterances in each test category

Statistic	Acronym
Arithmetic Mean	Mean
Mean divided by Standard Deviation	NormMean
5 percentile	Pct5
20 percentile	Pct20
30 percentile	Pct30
Mean below median	Mean50
Mean below 30 percentile	Mean30

Table 2. Utterance-level phone-posterior statistics that were studied

the QUOTES task. For the ALPHA test set, all the new statistics give better performance than the baseline arithmetic mean of the phone log posteriors. Of these, the normalized mean performed the best. For the QUOTES task, the normalized mean did not change the performance; however, the 5th percentile, and the mean below the median gave worse results than the baseline. This may be because there are far fewer phones per utterance in this case (as compared to the ALPHA task), and the corresponding estimates of the Pct5 and Mean50 are noisy.

We ran similar experiments and plotted ROC curves, for all the other test sets. For DIG16, the normalized mean gave an improvement, though smaller than that for ALPHA. For all other tasks, the normalized mean did not help or hurt. The other statistics were not useful by themselves either.

Next we explored whether a combination of the various statistics could improve performance using the approach described in Section 4. We trained Gaussian classifiers with feature vectors that included all the statistics in Table 2. To get the maximum possible gain for initial evaluation purposes, we trained the classifiers on the test data itself. The resulting likelihood ratio score (Equation 2) did not give any improvement over the normalized mean score. Similarly, Fischer's linear discriminant also did not improve over the normalized mean score.

Table 3 shows the equal error rate (EER) for the Mean and NormMean statistics. The largest improvements were attained using the NormMean statistic. We achieved a 26.7% relative improvement in the EER for the ALPHA test set, while maintaining performance for the other test sets.

EERs don't describe the shape of the curve away from the single point reported. They are appropriate when the number of IG utterances is equal to the number of OOG utterances. However, the most likely operating point for a practical application is not the EER point. From an application success point of view, a Total Error (TE) measure is more relevant. While each task has a different ratio of IG and OOG utterances, we computed TE by using a 85% IG to 15% OOG proportion across all test sets:

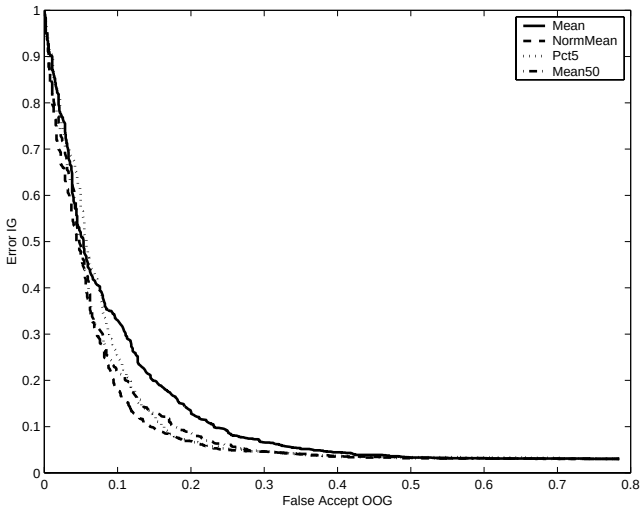

Fig. 1. ROC for ALPHA test set

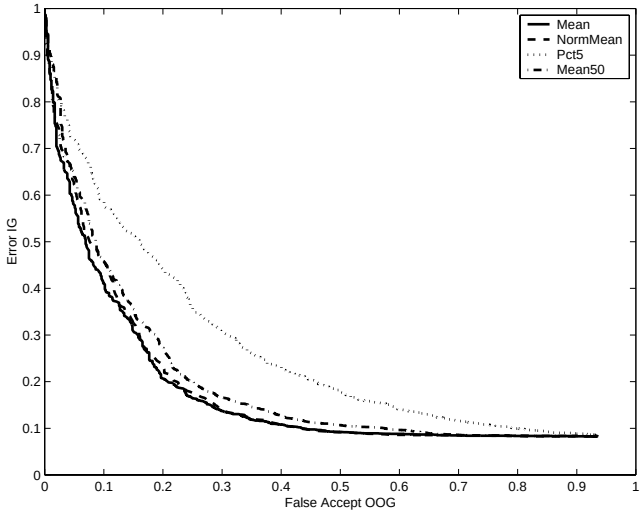

Fig. 2. ROC for QUOTES test set

$$TE = 0.85 * (FA_{IG} + FR_{IG}) + 0.15 * FA_{OOG}$$

Table 4 shows the best total error for the two test sets where the normalized mean showed an improvement. These are the ALPHA and the DIG16 test sets. The normalized mean had no significant positive or negative effect for the MENU, NAMES, and QUOTES tasks. The columns of the table are Error IG ($= FA_{IG} + FR_{IG}$), FA_{OOG}, and Total Error (TE).

The table shows a relative gain of 13% in FA_{OOG} at approximately the same IG accuracy for the DIG16 task. The relative difference in total error rate is small, about 2%. For the ALPHA task, the table shows a relative gain of 32% in FA_{OOG} at approximately the same IG accuracy. The relative difference in total error rate is 18%.

Data Type	EER	
	Mean	Normalized Mean (Relative diff.)
DIG16	11.7%	11.7% (0%)
ALPHA	17.2%	12.6% (26.7%)
MENU	12.8%	13.2% (-3.1%)
NAMES	23.7%	23.8% (-0.4%)
QUOTES	20.3%	21.1% (-3.9%)

Table 3. Equal Error Rates

Task	Confidence Score	Error IG	FA_{OOG}	Total Error (TE)
DIG16	Mean	11.6%	12.2%	11.7%
	Normalized Mean	11.9%	10.6%	11.5%
ALPHA	Mean	4.5%	39.2%	9.7%
	Normalized Mean	4.7%	26.5%	8.0%

Table 4. Best Total Error for two different confidence scores

6. SUMMARY AND CONCLUSION

We have presented several utterance-level statistics of phone posterior probability estimates for improving utterance rejection performance. We considered different ways of combining these statistics. In particular, we showed that a simple normalization of the mean by the standard deviation significantly improves the rejection performance for a fixed-length digit string task and a variable-length alpha-digit task. For a variety of other tasks we evaluated, the normalization technique had a negligible effect. Other combinations of the utterance-level statistics offered no advantage over the simple normalization scheme.

It is interesting to view these results in light of our original design of the confidence score. Our motivation in developing the normalized mean score was to improve rejection performance for cases where OOG utterances could be phonetically very similar to IG utterances. On examination of the tasks, we found that most of the OOG phrases for the alpha-digit task were "near" IG, i.e., they differed from an IG phrase in only one or two alphas or digits. In the case of the fixed-length digit strings, about a third of the OOG data contained digits, and the rest did not. It is in these two data sets that we observed the biggest wins, with a larger improvement for the alpha-digits task.

In the remaining three tests, the majority of the OOG phrases appear to be fairly disparate from the IG phrases. Our approach is not expected to offer a significant advantage over the mean posterior probability score for these cases. Again, the experiments bore out these expectations.

In summary, a simple normalization of the mean of the phone posterior probabilities by the standard deviation gives a significant improvement for the target tasks where OOG utterances can be highly confusable with IG utterances, while not affecting tasks where this confusion is not so great.

7. REFERENCES

[1] G. Williams and S. Renals, "Confidence measures for hybrid HMM/ANN speech recognition," in *Proceedings of Eurospeech-97*, 1997.

[2] F. Wessel, R. Schlüter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 288–298, 2001.

[3] M. Weintraub, F. Beaufays, Z. Rivlin, Y. Konig, and A. Stolcke, "Neural-network based measures of confidence for word recognition," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1997.

[4] Z. Rivlin, M. Cohen, V. Abrash, and T. Chung, "A Phone-Dependent Confidence Measure for Utterance Rejection," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 515–517, 1996.

[5] E. Chang, "Improving Rejection with Semantic Slot-Based Confidence Scores," in *Proceedings of EUROSPEECH*, pp. 271–274, 1999.

[6] M. G. Rahim, C.-H. Lee, and B.-H. Juang, "Robust utterance verification for connected digits recognition," in *Proceedings of ICASSP-95*, 1995.

[7] S. O. Kamppari and T. J. Hazen, "Word and phone level acoustic confidence scoring," in *Proceedings of ICASSP-00*, 2000.

[8] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*. John Wiley & Sons, 1973.

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