# IMPROVED UTTERANCE REJECTION USING LENGTH DEPENDENT THRESHOLDS

*Sunil K. Gupta and Frank K. Soong*

Bell Laboratories - Lucent Technologies, New Jersey, USA

## ABSTRACT

In this paper, we propose to use an utterance length (duration) dependent threshold for rejecting an unknown input utterance with a general speech (garbage) model. A general speech model, comparing with more sophisticated anti-subword models, is a more viable solution to the utterance rejection problem for low-cost applications with stringent storage and computational constraints. However, the rejection performance using such a general model with a fixed, universal rejection threshold is in general worse than the anti-models with higher discriminations. Without adding complexities to the rejection algorithm, we propose to vary the rejection threshold according to the utterance length. The experimental results show that significant improvement in rejection performance can be obtained by using the proposed, length dependent rejection threshold over a fixed threshold. We investigate utterance rejection in a command phrase recognition task. The equal error rate, a good figure of merit for calibrating the performance of utterance verification algorithms, is reduced by almost 23% when the proposed length dependent threshold is used.

## 1. INTRODUCTION

In this paper, we have investigated the problem of rejecting an unknown input utterance using a general speech model. The state of the art speech recognition systems, especially one operating in an "open-mic" mode generally need a rejection algorithm to accept or reject a recognized utterance. The rejection is typically formulated as a hypothesis testing procedure. In statistical hypothesis testing, the null hypothesis, $H_0$, that the input speech utterance $\mathbf{O} = \vec{o}_1, \vec{o}_2, \ldots, \vec{o}_T$ is correctly recognized, is tested against the alternate hypothesis, $H_1$, that the input utterance is incorrectly recognized. Note that alternate hypothesis includes cases where an in-grammar utterance is classified incorrectly as other in-grammar phrases and all out-of-grammar utterances. If the probability distribution for the null and alternative hypothesis are known exactly, then according to the Neyman-Pearson Lemma, the optimal test (in the sense of maximum power test) is the likelihood ratio test. The null hypothesis, $H_0$, is accepted if the likelihood ratio between the null and alternate hypothesis exceeds a critical threshold, rejected, otherwise[1]. This criterion expressed in log-domain and normalized by the utterance length is,

$$L_r = \frac{1}{T}(\log P(\mathbf{O}|H_0) - \log P(\mathbf{O}|H_1)) > \eta, \qquad (1)$$

where $T$ is the length of the input utterance, $\log P(\mathbf{O}|H_0)$ and $\log P(\mathbf{O}|H_1)$ are the log-probability of the input utterance for the null hypothesis and the alternate hypothesis, respectively, and $\eta$ is the *critical threshold* of the test.

Significant work has been done in the areas of keyword spotting and non-keyword rejection using general speech models. In [2], the likelihood of a filler (or garbage) model was used to construct a score for detecting keywords. In [3], a set of features including the likelihood of a garbage model were used to form a classifier for rejecting both non-keywords and recognition errors. Recently, anti-subword models have been used to perform utterance verification. In [4] a discriminatively trained, vocabulary independent utterance verification using anti-subword models was proposed. This method attempts to reject speech utterances contains no keywords and keywords but incorrectly recognized. In other work [5, 6], discriminative training procedure [7] was used to train anti-digit models to improve the rejection performance for connected digit recognition. In [6], techniques were proposed to adapt the rejection threshold to improve rejection performance in mismatched training and testing conditions. Rejection of a keyword is usually conducted either at segment or utterance level.

In this paper, we address the issue of improving rejection performance without using anti-models for rejection. The storage and/or computational complexity constraints of a "thin" DSP-based recognizer justify such an investigation. Typical consumer products like cellular phones, digital answering machines in which enhanced features like automatic speech recognition are highly desirable as long as only minimal cost is added. Under the above mentioned physical constraints, we investigate how to improve rejection performance using only a simple, general speech (garbage) model.

The probability of alternative hypothesis is computed based upon the general speech (garbage) model while the null hypothesis is evaluated using the phone or word models. We propose the use of a threshold $\eta$ which depends upon the utterance length. We show that for our recognition task, significant rejection performance improvement can be obtained, particularly for short utterances by using rejection threshold that varies with the length of input utterance. We investigate utterance rejection in a command phrase recognition task. A database of digit strings is used for rejecting out-of-vocabulary (OOV) utterances.

## 2. REJECTION USING UTTERANCE LENGTH DEPENDENT THRESHOLD

For our HMM-based speech recognition system, the log-probability of the input utterance given the null-hypothesis, $\log P(\mathbf{O}|H_0)$, is estimated as the log-likelihood of the recognized utterance for the decoding grammar. Figure 1 shows an example of a grammar which can be used to recognize one of $n$ different phrases.

The log-likelihood of the input utterance given the alternate hy-

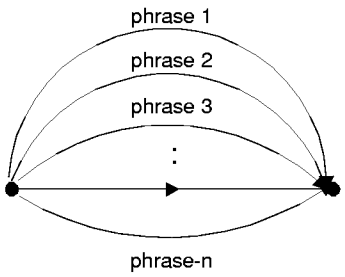


Figure 1: Phrase grammar.

pothesis, is computed over a three state hidden Markov model, trained on a large speech database. This model, known as a garbage model, represents the broad general characteristics of speech signals. Previously, single state general speech model was similarly used for speaker verification [8]. In our experiments, the three-state model (with same topology as sub-word models) are used as a generic phoneme or garbage model. we found in our experiments that the single-state model did not perform as well as the three-state garbage model. A grammar shown in figure 2 is used to compute the log-likelihood of the input utterance, given that it belongs to the alternate hypothesis.

Note that in Eq. 1, the utterance length is used as a normalization factor. The rejection criterion is then given by the following equation.

$$L_r \begin{cases} \geq \eta, & accept, \\ otherwise, & reject \end{cases} \tag{2}$$

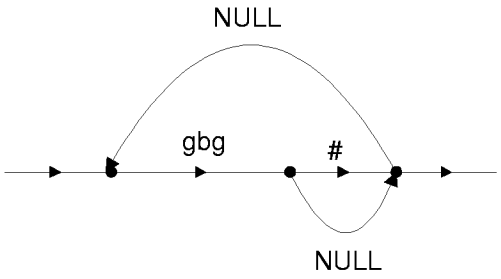


Figure 2: Garbage loop grammar used to obtain the likelihood of alternative hypothesis. Note that "gbg" represents the garbage model and "#" represents the background (silence) model.

In this paper, we have used the following modified likelihood ratio measure, normalized by the magnitude of the log-likelihood value $|\log P(\mathbf{O}|H_0)|$ and expressed as a percentage, as described below.

$$L_r = \frac{(\log P(\mathbf{O}|H_0) - \log P(\mathbf{O}|H_1))}{|\log P(\mathbf{O}|H_0)|} \cdot 100, \tag{3}$$

We find that the modified likelihood ratio measure in Eq. 3 tends to be more resilient to changes of grammar used in recognition.

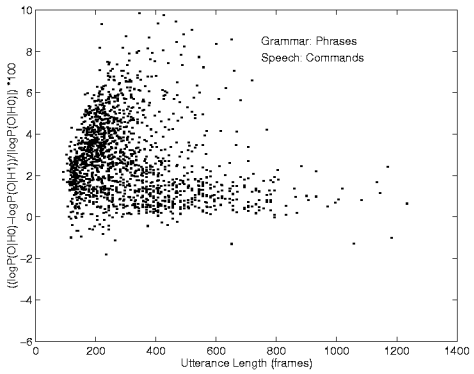


Figure 3: Scatter plot of the modified likelihood ratio for in-grammar command phrases. These include the correctly recognized as well as incorrectly recognized utterances.

The normalization used in the equation is also an implicit normalization by the utterance length. This confidence measure has been used in [9] for verbal information verification. Figure 3 shows a scatter plot of the modified likelihood ratio as a function of the utterance length, for a database of command phrases, given the command phrase grammar. The command phrases are used as in-grammar utterances.

Figure 4 shows a scatter plot of the likelihood ratios for out-of-vocabulary digit string utterances. Digit strings are used in this study as out-of-vocabulary (OOV) utterances. Several points can be noted from the two figures.

1. The likelihood ratio for in-grammar utterances is mostly positive for all utterance lengths. However, the likelihood ratio for out-of-vocabulary utterances, is negative for long utterances while a significant number of short utterances have positive likelihood ratios.
2. For long utterances, the distribution of likelihood ratios has smaller standard deviation than the standard deviation for short utterances.

When more flexible grammar such as a free-phone decoding is used, more alternative search paths are allowed and higher likelihood values can thus be obtained than a more rigid grammar. Also, while pruning techniques like beam search can become more effective when the decoding search is deep into the utterance, the phrase grammar of the null hypothesis presents very little constraint in short utterance decoding. The garbage loop grammar for the alternative hypothesis, on the other hand, imposes more or less uniform decoding constraints, independent of the utterance length. As a consequence of varying level of constraints for the null and alternate hypothesis, the likelihood ratio shows different distribution for different utterance lengths. As a result rejection threshold should be chosen as a function of utterance length to obtain a better performance. we propose to model the threshold $\eta$ as a polynomial function of the utterance length. That is, the threshold is

$$\eta(T) = a_{n-1}T^{n-1} + a_{n-2}T^{n-2} + \ldots + a_1T^1 + a_0, \tag{4}$$

where, $n-1$ is the order of the polynomial, and $a_i, 0 \leq i \leq n$ are the coefficients of the polynomial. Note that $n = 2$ results in a first-order, linear approximation. Further simplification leads to a piece-wise constant function. Let $S_T = T_i, 0 < i \leq N$ represent a set of utterance lengths such that

$$T \in T_i, \ if \ T_{i-1} \leq T < T_i, \tag{5}$$

$T_N = \infty$ and $T_0 = 0$. A separate rejection threshold $\eta_i, 0 \leq i < N$ is derived for each interval representing the input utterance lengths.

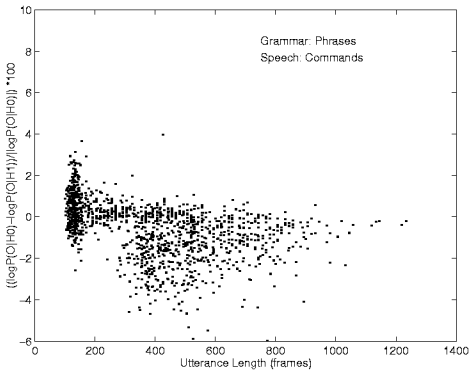


Figure 4: Scatter plot of modified likelihood ratio for a command-phrase grammar with digit strings as OOV utterances.

Next we describe our experimental setup and the resultant rejection performance using the proposed algorithm. We compare the rejection performance between a fixed and the proposed utterance length dependent thresholds.

## 3. EXPERIMENTAL RESULTS

In our experiments, we use mono-phone (i.e., context-independent), sub-word units. Each sub-word unit is a modeled as 3-state hidden Markov model with 8 Gaussian mixture components per state. Each digit is modeled as a 16-states HMM and each state is parameterized by 8 mixture Gaussian components. The background (silence) model is a single-state, 16-mixture component model. Once every 10 ms, twenty-five features (12-cepstral, 12-delta cepstral and 1 delta-energy) are computed for a frame of 30 ms speech samples. A separate 3-state, 64-mixtures per state, general speech (garbage) model is trained using digit strings and command phrases.

A test database of 1,638 phrases is used to perform the recognition test. The command phrases are in-grammar sentences. A 1,327 connected digit strings are used for testing out-of-vocabulary rejection. The digit database consists of strings of varying lengths (1, 4, 7, and 10 digits).

A real-time recognizer was used for recognition experiments using sequential Cepstral Mean Subtraction (CMS) to equalize possible channel difference between training and testing data. The baseline recognition accuracy for the phrase database is 90.5%. For rejection experiments, the misrecognized utterances (9.52%) were used in conjunction with the digit strings as alternative hypothesis (since the misrecognized utterances should also be rejected).

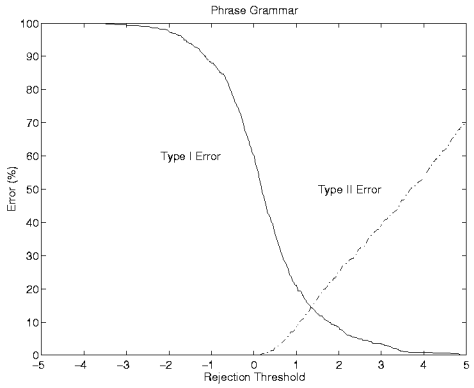


Figure 5: Receiver operating characteristic when a constant threshold is used for all utterance lengths.

Figure 5 shows the receiver operating characteristics (ROC) as a function of the rejection threshold when a single threshold is used to make an accept/reject decision.

In our experiment we partition the set of input utterances into several bins according to their utterance lengths. Both the command phrases and the digit strings are divided into two approximately equal sets. The first set of sentences from the phrase and digit databases are used to derive rejection thresholds for achieving equal error rate. The equal error rate rejection thresholds so derived are then used to evaluate the performance on the second set of sentences. Table 1 shows the Type I and Type II errors when a single, fixed threshold is used for all utterances. Table 2 shows the errors when different rejection thresholds are selected according to the interval to which the utterance length belongs. Several remarks can be made:

1. Short utterances contribute towards majority of the overall errors.
2. A comparison of the first rows in table 1 and table 2 shows that there is a reduction of errors for short utterances (less than 200 frames) by almost 30%. The error reduction for very long utterances (longer than 450 frames) is about 26%. The performance for utterances with intermediate lengths (between 200 and 450 frames) is basically unchanged.
3. Table 2 shows that lower rejection thresholds should be used for longer utterances. This is consistent with the proposed algorithm in this paper that the rejection threshold should be modeled as a length dependent variable, rather than a constant.
4. There is an overall improvement of 22.5% in rejection performance.

Figure 6 shows the equal error rate threshold as a function of the utterance length. Note that the equal error rate threshold for shorter utterances is significantly higher than the fixed, sin-

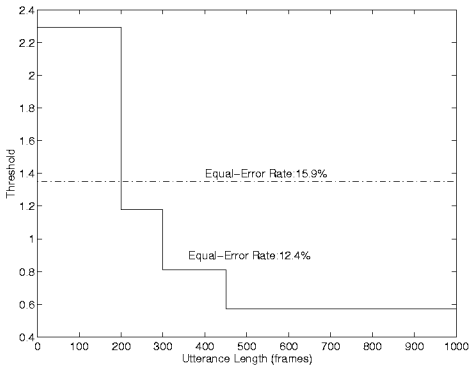


Figure 6: A comparison of the equal error rate threshold when a single threshold is selected with the variable equal error rate threshold for different utterance length intervals.

gle threshold. For longer utterances, the length dependent equal error rate threshold is lower than the fixed threshold.

It is important to point out that a different approximation to the polynomial of equation 4 (other than the piecewise constant approximation used in our experiments) could be employed to further enhance the performance. In particular, a first-ordered, linear approximation should be particularly beneficial for shorter utterances.

| Length (in frames) | Threshold | # in Error/Total Utterances | | |
|---|---|---|---|---|
| | | Type I | Type II | Total |
| **0-200** | **1.35** | **21 / 273** | **107 / 256** | **128 / 529** |
| 200-300 | 1.35 | 20 / 212 | 4 / 65 | 24 / 277 |
| 300-450 | 1.35 | 35 / 145 | 5 / 188 | 30 / 333 |
| $\geq$450 | 1.35 | 42/ 111 | 3 / 232 | 45 / 333 |
| **Total** | **1.35** | **118 / 741** | **119 / 741** | **237 / 1482** |

Table 1: Rejection performance on test database for different utterance lengths for single equal error rate rejection threshold.

# 4. DISCUSSION

In this paper, we have presented a simple but effective technique to improve the rejection performance of an automatic speech recognition system. A general speech model is used to obtain the likelihood for the alternative hypothesis in making an accept/reject decision. We have shown that the normalized likelihood ratio can be modeled as a high-order polynomial in utterance length than just a fixed constant. The experiments show that the a simple, piecewise constant approximation to the polynomial results in a reduction of almost 23% in both false alarms (Type II error) and false rejection (Type I) errors at equal error rate. The rejection performance of shorter utterances can be improved significantly when an utterance length dependent rejection threshold is used.

The technique presented in this paper can be applied to improve the performance of barge-in, or detecting a partial valid phrase before the end of an utterance. Just like rejection, detection of a partial valid phrase is more difficult in the earlier stage than later. A length-dependent threshold should be equally effective in improving the barge-in detection, especially in the early part of the utterance.

| Length (in frames) | Threshold | # in Error/Total Utterances | | |
|---|---|---|---|---|
| | | Type I | Type II | Total |
| **0-200** | **2.29** | **46 / 273** | **44 / 256** | **90 / 529** |
| 200-300 | 1.18 | 16 / 212 | 5 / 65 | 19 / 277 |
| 300-450 | 0.81 | 18 / 145 | 22 / 188 | 40 / 333 |
| $\geq$450 | 0.57 | 11/ 111 | 22 / 232 | 33 / 333 |
| **Total** | | **91 / 741** | **93 / 741** | **184 / 1482** |

Table 2: Rejection performance on test database for different utterance lengths for a different equal error rate rejection threshold for each utterance length interval.

The length dependent threshold if incorporated explicitly in a rejection algorithm where more sophisticated, discriminatively trained anti-models are used, the high performance can be even further improved.

# 5. REFERENCES


1. K. Fukunaga, *Statistical Pattern Recognition*. Academic Press Inc., second ed., 1990.
2. R. C. Rose, "Discriminant word spotting techniques for rejecting non-vocabulary utterances in unconstrained speech," in *Proc. Int. Conf. Acoust. Speech Sign. Process.*, vol. 2, pp. 105–108, 1992.
3. B. Chigier, "Rejection and keyword spotting algorithms for a directory assistance city name recognition application," in *Proc. Int. Conf. Acoust. Speech Sign. Process.*, vol. 2, pp. 93–96, 1992.
4. R. A. Sukkar and C.-H. Lee, "Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 6, pp. 420–429, 1996.
5. M. G. Rahim, C.-H. Lee, and B.-H. Juang, "Discriminative utterance verification for connected digit recognition," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, pp. 266–277, 1997.
6. M. G. Rahim, C.-H. Lee, and B.-H. Juang, "A study on robust utterance verification for connected digit recognition," *J. Acoust. Soc. Am.*, vol. 101, no. 5, pp. 2892–2902, 1997.
7. B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Transactions on Signal Processing*, vol. 40, pp. 3043–305, 1992.
8. S. Parthasarathy and A. E. Rosenberg, "General phrase speaker verification using sub-word background models and likelihood ratio scoring," in *Proc. Int. Conf. on Spoken Language Processing.*, vol. 4, pp. 2403–2406, 1996.
9. Q. Li, B.-H. Juang, Q. Zhou, and C.-H. Lee, "Verbal information verification," in *Proc. European Conference on Speech Communication and Technology.*, vol. 2, pp. 839–842, 1997.