Acoustic Modeling for the SRI Hub4 Partitioned Evaluation Continuous Speech Recognition System

Ananth Sankar, Larry Heck, and Andreas Stolcke

Speech Technology And Research Laboratory
SRI International
Menlo Park, California

ABSTRACT

We describe the development of the SRI system evaluated in the 1996 DARPA continuous speech recognition (CSR) Hub4 partitioned evaluation (PE). The task for the Hub4 evaluation was to recognize speech from broadcast television and radio shows. Recognizing such speech by machines poses many challenges. First, the segments to be recognized could be very long. This introduces a problem in training and recognition because of the consequent increased system memory requirement. A simple segmentation technique is used to break long segments into shorter, more manageable lengths. The speech from broadcast news sources exhibits a variety of difficult acoustic conditions, such as spontaneous speech, band-limited speech, and speech in the presence of noise, music, or background speakers. Such background conditions lead to significant degradation in performance. We describe techniques, based on acoustic adaptation, that adapt recognition models to the different acoustic background conditions, so as to improve recognition performance. We also present a novel algorithm that clusters the test data segments so that the resulting clusters are homogeneous with respect to speakers. This is followed by acoustic adaptation to the individual clusters, resulting in a significant performance improvement. Finally, we briefly describe our studies in language modeling for the Hub4 evaluation which is detailed further in another paper in these proceedings.

1. Introduction

The test paradigm for the 1996 DARPA-sponsored Hub4 continuous speech recognition (CSR) evaluation was broadcast television and radio speech. Speech from these sources is natural, and exhibits a variety of qualities that, taken as a whole, makes recognizing it by machine an interesting challenge. First, the segments to be recognized are sometimes very long (on the order of a few minutes), making the training and recognition task harder because of increased memory burdens on the system. Second, the speech exhibits a variety of acoustic background conditions, such as degraded or music-corrupted speech, and different speaking styles, such as planned or spontaneous. It is hoped that research directed toward this task will result in robust speech recognition systems that will perform well across a variety of acoustic conditions.

In this paper, we present the system developed at SRI for the 1996 Hub4 PE. In Section 2, we briefly describe the Hub4 partitioned evaluation (PE) task, and present the individual components of the SRI system (see also Figure 1), which are detailed in later sections. In Section 3, we present an algorithm to automatically segment long sentences into nominally 10 second segments for further processing. This is necessary in order to minimize the memory burden caused by very long data segments. In Section 4, we describe the application of our previously developed acoustic adaptation algorithms to train recognition models that are specific to different acoustic focus conditions observed in the data. These condition-specific models are then used to recognize test segments that are classified as belonging to those acoustic conditions. In Section 5, we present a novel clustering and adaptation algorithm that is used to adapt the condition-specific models to the test data. The idea is to cluster the test segments into speaker-homogeneous clusters and then to adapt the recognition models to these clusters. These models are then used to recognize the data in these clusters. In Section 7, we briefly describe our language model (LM) techniques for this evaluation. Our language modeling work is presented in greater detail in another paper in these proceedings [1]. We conclude in Section 8 with a summary of our work.

2. A high-level description of the SRI Hub4 PE system

For 1996, the Hub4 evaluation was divided into two individual problems. In the unpartitioned evaluation (UE), the test data consisted of a set of television and radio shows in their entirety. However, commercials and sports reports were not included in the data to be recognized, as these were considered to be very different in type of language as compared to the rest of the data. In the UE, it is necessary to automatically excise the speech segments from the test data before recognizing them. In the PE, each test show was partitioned into segments of speech. Thus, pure music or noise segments were removed by hand, and only speech segments remained. Each segment contained speech from a single speaker. In addition, the segments were homogeneous with respect to the acoustic background condition or speech style. The segments were classified into seven different acoustic focus conditions, F0, F1, F2, F3, F4, F5, FX, as described in [2], and the labels were provided for use in the evaluation.

The SRI Hub4 PE system used in the 1996 DARPA CSR evaluations involved the following stages in processing the recognition data:

1. The long segments in each of the focus conditions was broken into nominally 10 second segments by using an automatic segmentation algorithm (see Section 3). The segments were then processed differently, depending on the acoustic focus condition.

2. The front-end feature extraction was based on mel-frequency cepstrum processing. The original speech data was sampled at 16,000 samples per second. For the F2 (telephone) segments, the speech was bandlimited, and downsampled to 8,000 samples per second. To extract features, the speech was then hamming-windowed with a 25.6 ms window, and the window was advanced every 10 ms. Each frame was represented by 12 mel-frequency cepstrum coefficients, the log energy, and their first- and second-order time derivatives (delta and delta-delta features), for a resulting 39-dimensional feature vector.

3. For each focus condition, we generated word lattices for each segment by using the lattice generation algorithm described in [3]. We used condition-specific acoustic models estimated with the training data for each focus condition, using maximum-likelihood transformation-based adaptation techniques [4, 5, 6]. The condition-specific models were estimated by adapting seed models that were trained using either the Wall Street Journal (WSJ) SI-284 database, or the Switchboard and Macrophone databases (see Section 4). For this stage, we used a 20,000-word bigram LM trained on the Hub4 LM training data and the transcripts of the Hub4 acoustic training data. The condition-adapted acoustic models and the bigram LMs were used to generate recognition hypotheses for each test segment with these lattices. These recognition hypotheses were later used to adapt the condition-adapted models to the test conditions.

4. The test segments for each focus condition were clustered by using an agglomerative clustering algorithm [7]. The condition-adapted models were then separately adapted to each test cluster (see Section 5).

5. N-best lists were generated using the models adapted to each test cluster. These lists were then rescored with larger trigram and fourgram LMs [1] as described in Section 6.

Figure 1 schematically describes the system for the F0 test data. The data from other focus conditions were similarly processed, except that the seed models used for adaptation to the F2 data were trained on Switchboard and Macrophone databases.

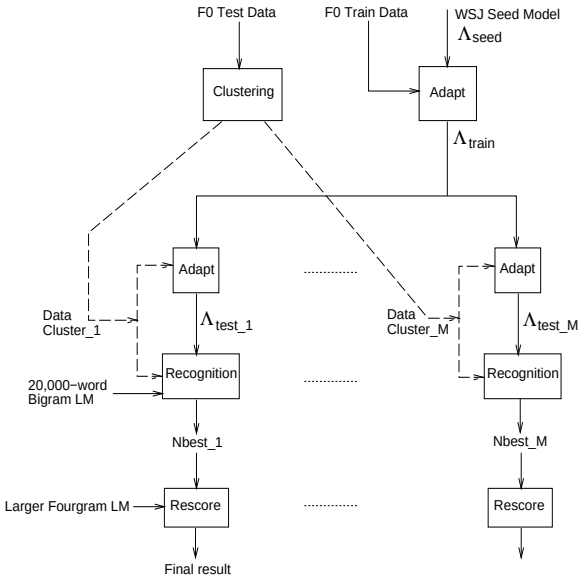

Figure 1: Schematic of the SRI Hub4 PE system

3. Automatic Segmentation of Long Segments

LDC transcribed about 35 hours of acoustic material collected from different television and radio shows to serve as training material for the evaluations. These segments, as well as the segments to be recognized in the H4-PE evaluation, were annotated as being in one of the seven acoustic focus conditions.

The acoustic segments derived from the show episodes varied in length. Many of the segments were quite long. Since handling longer segments placed an increased memory requirement on our system, we further segmented these long segments into smaller segments, which were nominally 10 seconds in duration. This was accomplished by using a gender-independent phonetically tied mixture (PTM) hidden Markov model (HMM) system with low pruning thresholds to run a continuous backtrace search on the long segments. If all hypotheses pass through a grammar node at the same time, then the continuous backtrace search would output the hypothesis at that node and reset the backtrace memory. The result of this step was a hypothesis along with the word-level backtrace information giving word start and end times.

This backtrace was then processed by looking for silence or pause regions that fall at nominally 10 second intervals. Such silence or pause regions were marked as sentence boundaries. We then broke up the long segments at these sentence boundaries to create nominally 10-second segments for further processing.

During training we used a slight variation of this procedure for breaking up the segments. Since we have access to the correct transcriptions for the training data, we used this procedure

but instead of running a continuous backtrace recognizer to get the word-level backtrace information, we simply force-aligned the correct transcription against the PTM models to produce a word-level backtrace from which pause information was extracted to break up the segments. These segments along with their corresponding transcriptions were used for training. Since the original transcribed data contained some annotation and transcription errors, we further verified the training data segments by looking for obvious inconsistencies between the acoustics and the transcripts. As a result of this procedure, we eliminated about 5% of the data.

4. Adaptation during training

Since only 35 hours of transcribed training data was available for all the focus conditions, we decided that the best strategy to train models for each condition would be to use maximum-likelihood (ML) transformation-based adaptation techniques [4, 5, 6] to adapt seed models to each condition. We have previously developed algorithms for ML transformation-based acoustic adaptation [4, 5, 6]. The general idea is to transform the test domain acoustic features or the trained HMM parameters to reduce the mismatch between them. The parameters of the transformation are estimated by maximizing the likelihood of adaptation data available from the test domain. We have developed techniques to adapt both the HMM means using a block-diagonal affine transformation [6] and the variances using a variance scaling transform [5, 6]. Other feature- and model-space transformations have also been detailed in our earlier work [4, 5, 6]. Full matrix affine transformations of the HMM mean vector have been previously studied [8]. However, we found that the block-diagonal approach was more robust. In this approach, a separate matrix affine transform is used to transform the cepstrum, delta cepstrum, and delta-delta cepstrum. Thus, a particular Gaussian mean vector $\boldsymbol{\mu}$ is transformed according to

$$\boldsymbol{\mu}'_{\text{cep}} = \boldsymbol{A}_{\text{cep}}\boldsymbol{\mu}_{\text{cep}} + \boldsymbol{b}_{\text{cep}}, \quad (1)$$

$$\boldsymbol{\mu}'_{\Delta\text{cep}} = \boldsymbol{A}_{\Delta\text{cep}}\boldsymbol{\mu}_{\Delta\text{cep}} + \boldsymbol{b}_{\Delta\text{cep}}, \quad (2)$$

$$\boldsymbol{\mu}'_{\Delta^2\text{cep}} = \boldsymbol{A}_{\Delta^2\text{cep}}\boldsymbol{\mu}_{\Delta^2\text{cep}} + \boldsymbol{b}_{\Delta^2\text{cep}}, \quad (3)$$

where $\boldsymbol{\mu}'$ is the transformed mean vector, $\boldsymbol{\mu}$ is the original mean vector, $\boldsymbol{A}$ is a transformation matrix, and $\boldsymbol{b}$ is a bias vector. The subscripts refer to the cepstrum, delta and delta-delta cepstrum features. Since a separate transformation is used for the cepstrum, delta cepstrum, and delta-delta cepstrum, this is a block-diagonal transformation of the full feature vector. The number of parameters that need to be estimated is much less than in a full matrix transformation, and is hence more robust. Since we use diagonal covariance matrices, the variance scaling transform scales each component of the variance vector by a scale factor,

$$\sigma^{2'} = \alpha\sigma^2, \quad (4)$$

where $\sigma^{2'}$ and σ^2 refer to a component of the transformed and untransformed variance vector, respectively, and α is the scale factor to be estimated.

As described in previous work [4, 5, 6], the parameters of the transformations are estimated by maximizing the likelihood of adaptation data from the new acoustic environment. Separate transforms are used for different Gaussian clusters as in [4], including a separate transform for the Gaussians corresponding to the silence model. We have made use of the transforms in Equations 1 through 3, and 4, for the Hub4 PE evaluation system.

Condition-specific HMMs were trained by adapting seed models to the individual acoustic focus conditions by using the 35 hours of transcribed Hub4 training data provided by LDC. For the condition-specific adapted models, we used only the block-diagonal mean adaptation technique. The seed models were gender-dependent and trained using Switchboard and Macrophone data for the F2 focus condition, and using the Wall Street Journal WSJ SI-284 database for all the other focus conditions. This choice was based on the results of an initial recognition experiment using both sets of models for all the focus conditions.

We used a portion of the 1996 H4 development test data to run recognition experiments, with the lattices generated with the condition-specific HMMs and the 20,000-word bigram LM. This portion of the development data corresponded to all segments that were originally longer than 10 seconds and had to be broken up by the acoustic segmentation algorithm described in Section 3.

Table 1 shows the word-error rates obtained for each of the focus conditions by using the seed models and the condition-adapted models. In the table, NT stands for the number of transforms used. As can be seen, a 9.9% relative improvement over the seed model was achieved with our adaptation techniques. For further processing, we chose to use 11 transforms, which gave slightly better performance than 3 transforms, as shown by the table.

5. Test-cluster-based adaptation

Condition-specific models are estimated using adaptation algorithms and the training data for each focus condition. However, there may still be a mismatch between these condition-specific models and test data from the same acoustic condition. Such mismatches are largely due to different speakers between training and testing. In addition, there may be small differences in the training and test acoustical conditions, leading to a mismatch. Since the main source of variability between the training and test conditions is the different speakers, we used an unsupervised bottom-up agglomerative clustering algorithm to cluster acoustic segments that were similar to each

Condition	Seed Models	Condition-adapted models	
		NT = 3	NT = 11
F0	25.7	22.2	22.1
F1	46.2	41.0	41.1
F2	47.8	46.9	46.9
F3	49.4	44.9	43.7
F4	44.9	37.0	36.4
F5	37.2	35.7	34.5
FX	68.8	62.9	63.0
All	44.5	40.3	40.1

Table 1: Performance of condition-adapted models

other. Since acoustic segments of the same speaker are similar, the resulting clusters are homogeneous with respect to speakers.

Once the segments are clustered, the condition-specific models are separately adapted to each cluster by using the block-diagonal mean transformation (Equation 1 through 3), followed by the variance scaling transformation (Equation 4). In this stage we used three separate transformations, including a separate transformation for the silence Gaussians. The reference transcriptions for adaptation were derived by running a one-pass Viterbi recognition search through the lattices with the condition-specific models used to generate the lattices. Once the models are adapted, it is possible to re-recognize the acoustic segments for each cluster and then re-adapt the models by using these new hypotheses. However, we did not observe a significant improvement with multiple iterations of this kind and hence we used only one iteration.

For clustering, the distance between two acoustic segments $\boldsymbol{X}_i = \{\boldsymbol{x}_{i,1}, \ldots, \boldsymbol{x}_{i,T_i}\}$ and $\boldsymbol{X}_j = \{\boldsymbol{x}_{j,1}, \ldots, \boldsymbol{x}_{j,T_j}\}$ was computed using a symmetric relative entropy distance,

$$D(i,j) = \frac{1}{T_i}\sum_{t=1}^{T_i} \log \frac{p(\boldsymbol{x}_{i,t} \mid \lambda_i)}{p(\boldsymbol{x}_{i,t} \mid \lambda_j)} + \frac{1}{T_j}\sum_{t=1}^{T_j} \log \frac{p(\boldsymbol{x}_{j,t} \mid \lambda_j)}{p(\boldsymbol{x}_{j,t} \mid \lambda_i)}, \quad (5)$$

where Λ_i and Λ_j are the underlying statistical models of X_i and X_j. The distance between two clusters was then computed as the maximum distance between segments in the two clusters [9]. In our work, we used a Gaussian mixture model (GMM) to model each test segment. This procedure was previously described by us in [7], but applied to cluster the training data speakers. In the work reported here, we used it to cluster the test data segments.

Since a mixture model must be trained for each segment to compute the relative entropy measure, and many of the segments were short in duration (some less than 1 second), we varied the number of Gaussians in the model of each segment based on a heuristic function of the segment duration

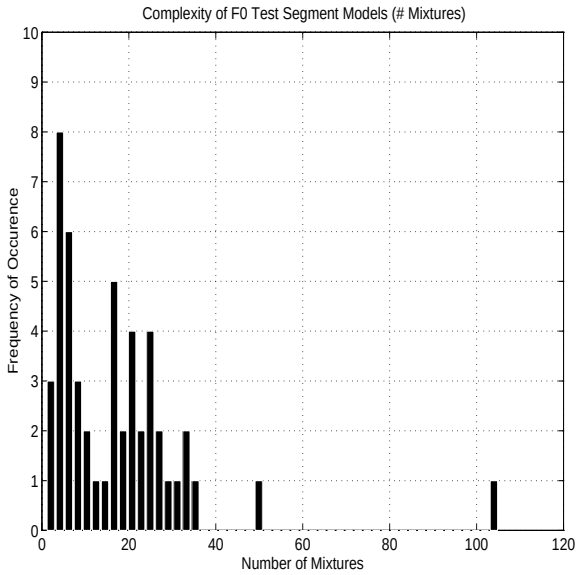

Figure 2: GMM sizes used to model the test segments

length. This prevented over-fitting of the model to the short data lengths. Figure 2 shows a histogram of the number of Gaussians used in the GMMs for each development segment of the F0 condition. As can be seen, the number of Gaussians varies from one to more than one-hundred.

A threshold on the minimum distance between any pair of clusters defines a cut in the agglomerative cluster tree and hence a set of test segment clusters. This threshold was empirically determined. By examining the clusters on the 1996 H4 development set, we found that the clusters were indeed quite homogeneous with respect to speakers.

Table 2 shows the improvement in performance obtained by adapting the condition-specific models to the test data clusters. The table shows the word-error rates with the condition-specific models described in Section 4 and the models adapted to the test conditions. As in Table 1, recognition was performed using the word lattices and a 20,000-word bigram LM. The condition-specific error rates in Tables 1 and 2 do not match because we used the entire development test set for the experiments reported in Table 2, whereas only the segments longer than 10 seconds were used for the previous experiment. We achieved a significant 9% error-rate improvement by using test-cluster-based adaptation.

To evaluate the advantage of performing the unsupervised speaker clustering, we also adapted the condition-specific models to each test condition without doing any clustering. Thus, in this case, all the data from any test condition was used to adapt the models as opposed to only the data in each of the speaker clusters in the test data. Since using all the test data allowed us to estimate a larger number of transfor-

Condition	Models		
	Condition-specific	Test-cluster-adapted	
		Mean	Mean and variance
F0	22.6	21.3	20.8
F1	41.2	38.8	38.8
F2	47.2	44.0	42.3
F3	45.6	42.5	42.4
F4	36.9	34.4	33.8
F5	36.3	28.3	28.1
FX	63.8	57.8	57.2
All	41.2	37.8	37.3

Table 2: Performance of test-condition-adapted models

mations, we used 11 transformations, including a separate transformation for the silence Gaussians, as compared to 3 transformations in the case of adapting to the speaker clusters. Table 3 shows the advantage of using the unsupervised clustering method over simply adapting to the test conditions. The second two columns show the word-error rate when the condition-specific models were adapted to the test conditions, using all the data in each acoustic focus condition (single cluster). The last two columns are replicated from Table 2 and show the performance after adaptation to the test data clusters (multiple clusters). We can see that adapting to the individual test conditions gave a relative improvement of 3.4% as compared to the condition-specific models, and adapting to the test data clusters gave a further 6.0% improvement, resulting in a total relative improvement of 9.2% compared to the condition-specific models. It is clear that adapting to the test data clusters gave a consistent improvement compared to adapting to only the test conditions for all acoustic conditions.

Condition	Adapt to test conditions			
	Single Cluster		Multiple clusters	
	Mean	Mean and variance	Mean	Mean and variance
F0	22.4	22.7	21.3	20.8
F1	40.0	40.3	38.8	38.8
F2	46.9	46.2	44.0	42.3
F3	44.7	44.8	42.5	42.4
F4	35.0	34.5	34.4	33.8
F5	31.4	31.1	28.3	28.1
FX	62.2	62.2	57.8	57.2
All	39.9	39.8	37.8	37.4

Table 3: Effect of clustering

Vocabulary Size (words)	Word-error rate (%)
20,000	33.3
48,000	32.7

Table 4: Effect of vocabulary size

6. Rescoring of N-best lists

The test-cluster-adapted models were used to generate N-best lists from the word lattices for all segments in a test cluster. The 20,000-word bigram LMs were used to generate the N-best lists. These lists were then rescored using the test-cluster-based acoustic models, word-transition penalty, and a larger trigram/fourgram interpolated LM, which is described in Section 7. The different knowledge source scores were combined using linear weighting, where the weights were estimated with a grid search to optimize the error rate on the development test data. We estimated three sets of weights for the acoustic condition groups F0, F4, and F5; F1, F2, and F3; and FX. These groups were chosen because the word-error rates were similar within them.

The word-error rate with the test-cluster-adapted models and bigram LMs was 37.0%. When the N-best lists were rescored with the same acoustic models but with larger trigram and fourgram interpolated LMs [1], the error rate was 33.1%. This was the best performance we achieved on our development test set. Our error rate on the evaluation test data with this system was 33.3%. In the interest of time, we had decided not to use cross-word acoustic models and also to use a smaller vocabulary than usual. Our 20,000-word vocabulary resulted in a 2.1% out-of-vocabulary rate on the 1996 H4 development test data. We subsequently ran experiments with a 48,000-word vocabulary, which resulted in an out-of-vocabulary rate of 0.9% on the development data. This system gave a 32.7% word-error rate on the evaluation test set. This is summarized in Table 4. Thus, the increase in the vocabulary size gave a small improvement over the system we used for the evaluation.

7. Language models

The lattices used by our system were generated by a 20,000-word vocabulary, bigram back-off [10] LM trained using the 1996 H4 LM training texts and the transcripts for the H4 acoustic training data provided by LDC and NIST. The vocabulary for the LM was selected by choosing the most frequent words from the H4 LM training texts and adding all words that occurred at least twice in the acoustic training transcripts. A separate LM was trained with the H4 LM texts and the H4 acoustic training transcripts, and the conditional probabilities were linearly interpolated. The interpolation weight was roughly optimized to minimize the perplexity on the development test data.

The N-best lists were rescored with a larger LM based on four text sources: the Hub4 LM training texts, the Hub4 acoustic training transcripts, the NABN 1995 training data, and the Switchboard corpus training data. A fourgram interpolated language model was trained using these different databases.

Our language modeling studies for this evaluation are described in detail in another paper in these proceedings [1]. That paper also describes our studies of techniques to adapt the LMs to the acoustic focus conditions (since speaking styles could be correlated with these conditions), and to different topics.

8. Summary and conclusions

We have described the SRI system for the 1996 DARPA Hub4 PE. It was found that ML transformation-based adaptation of seed WSJ or Switchboard and Macrophone models to the acoustic focus conditions using the 35 hours of transcribed Hub4 data available from LDC provided a significant 9.9% improvement in performance. We presented a novel algorithm for adaptation during testing, which used an unsupervised agglomerative clustering algorithm to cluster the test segments, followed by ML transformation-based adaptation of the condition-specific models to these clusters. A symmetric relative entropy distance between test segments was used for clustering. We described a robust method to estimate the models for each test segment necessary for the computation of the distance measure. It was shown that adapting to all the test data in each focus condition gave a 3.4% decrease in error rate as compared to the condition-specific models. However, adapting to the individual clusters proved to be even more important and gave a 9% improvement over the condition-specific models. This paper focused mainly on the acoustic modeling components of the system. Our language modeling work is described in detail in another paper in these proceedings [1].

References

1. F. Weng, A. Stolcke, and A. Sankar, "Hub-4 Language Modeling using Domain Interpolation and Data Clustering," in *Proceedings of the DARPA Speech Recognition Workshop*, (Chantilly, VA), 1997.
2. R. Stern, "Specification of the 1996 Hub4 Broadcast News Evaluation," in *Proceedings of the DARPA Speech Recognition Workshop*, (Chantilly, VA), 1997.
3. H. Murveit, J. Butzberger, V. Digalakis, and M. Weintraub, "Large-Vocabulary Dictation Using SRI's DECIPHER(TM) Speech Recognition System: Progressive-Search Techniques," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. II–319–II–322, 1993.
4. V. Digalakis, D. Rtischev, and L. Neumeyer, "Speaker Adaptation Using Constrained Reestimation of Gaussian Mixtures," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 5, pp. 357–366, 1995.
5. A. Sankar and C.-H. Lee, "A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 190–202, May 1996.
6. L. Neumeyer, A. Sankar, and V. Digalakis, "A Comparative Study of Speaker Adaptation Techniques," in *Proceedings of EUROSPEECH*, pp. 1127–1130, 1995.
7. A. Sankar, F. Beaufays, and V. Digalakis, "Training Data Clustering for Improved Speech Recognition," in *Proceedings of EUROSPEECH*, 1995.
8. C. J. Legetter and P. C. Woodland, "Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression," in *Proceedings of the Spoken Language Systems Technology Workshop*, pp. 110–115, 1995.
9. R. Duda and P. Hart, *Pattern Classification and Scene Analysis*. John Wiley & Sons, 1973.
10. S. M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 3, pp. 400–401, 1987.