Analysis of Tradeoffs in Modeling Continuous Speech Recognition for Domain Specific Training Application

Laurie D. Marshall
NAVAIR Orlando
Orlando, Florida
laurie.marshall@navy.mil

Dave Ribakoff
NAVAIR Orlando
Orlando, Florida
david.ribakoff@navy.mil

Lisa Ouakil
NAVAIR Orlando
Orlando, Florida
lisa.ouakil@navy.mil

Mark Varvak
NAVAIR Orlando
Orlando, Florida
mark.varvak@navy.mil

Keywords:
speech recognition, speech technology, acoustic model, language model, Hidden Markov Model (HMM), Navy Air Traffic Control (ATC), perplexity, entropy, word accuracy

ABSTRACT

Since the 1980's the underlying technology in speech recognition has been the Hidden Markov Model (HMM), an accurate process to statistically model continuous human speech. A speech model is represented as combination of probabilities associated with both acoustic and language models. Acoustic models estimate the probability associated with postulated sequence of acoustic observations. Language models describe the probability associated with postulated sequence of words and can incorporate both syntactic and semantic constraints of the language. When developing speech recognition for training systems, both acoustic and language models are crafted for the application. Due to the complexity in building a tuned accurate speech recognition application, it is necessary to understand how acoustic and language models affect accuracy. The Speech Technology Group (STG) at NAVAIR Orlando develops acoustic and language models specifically for the Navy Air Traffic Control (ATC) trainers, in contrast to commercial-off-the-shelf speech tools that contain generic acoustic models with limited alterability. The present study evaluates several speech model configurations including word pair (bi-gram) models. The STG, under laboratory conditions, measured the effects of accuracy of the following variables: vocabulary, perplexity, acoustic models, and language models. The findings of this study describe the influence of acoustic and language modeling on speech recognition. These lessons learned provide a better understanding of how speech model parameters influence model accuracy and can be used to more efficiently incorporate speech recognition within training applications, thereby enhancing the learning performance of the war-fighter.

ABOUT THE AUTHORS

Laurie D. Marshall is an engineer in the Modeling and Simulation Development branch at NAVAIR Orlando Training Systems Division where she currently leads the Speech Technology Group in computational linguistic research and development efforts. She received a B.S. in Electrical Engineering with a minor in Applied Mathematics from University of Central Florida and a M.S. in Mechanical Engineering from Georgia Institute of Technology. She currently is pursuing a Doctorate degree in Modeling and Simulation.

Dave Ribakoff is a senior engineer within the Modeling and Simulation Development Branch at NAVAIR Orlando Training Systems Division where he specializes in developing acoustic models for the Speech Technology Group for military Air Traffic Control training applications. He holds a B.S. degree in Aerospace Engineering from Rensselaer Polytechnic Institute and a M.S. degree in Aerospace Engineering from University of Michigan.

Lisa Ouakil is an engineer within the Modeling and Simulation Development Branch at NAVAIR Orlando Training Systems Division where she conducts research in pattern recognition for the Speech Technology Group. She holds a B.S. degree in Computer Engineering from the University of Central Florida with a minor in Applied Mathematics and is currently in pursuit of an M.S. degree in Industrial Mathematics.

Mark Varvak is a Computer Engineer at NAVAIR Orlando. He acquired his B.S. in Applied Mathematics at State University of Ukraine and M.S. in Applied Mathematics at the University of Central Florida.

Analysis of Tradeoffs in Modeling Continuous Speech Recognition for Domain Specific Training Application

Laurie D. Marshall
NAVAIR Orlando
Orlando, Florida
laurie.marshall@navy.mil

Dave Ribakoff
NAVAIR Orlando
Orlando, Florida
david.ribakoff@navy.mil

Lisa Ouakil
NAVAIR Orlando
Orlando, Florida
lisa.ouakil@navy.mil

Mark Varvak
NAVAIR Orlando
Orlando, Florida
mark.varvak@navy.mil

INTRODUCTION

This article analyzes the influence of acoustic and language models within a speech recognition system. Figure 1 presents a diagram that illustrates one approach to implementing a speech recognition system. The Entropic Speech Recognition System of Cambridge Research Laboratory [8] is a speech modeling toolkit that implements this approach. The Speech Technology Group (STG) at NAVAIR Orlando has utilized and applied it to developing speech models for Navy Air Traffic Control (ATC) training systems. This speech recognition system was selected because it allows for the customization of acoustic models within a speech application. For comparison, the Nuance speech modeling toolkit was used because it contains an efficient generic acoustic model and has demonstrated a flexible approach to language modeling.

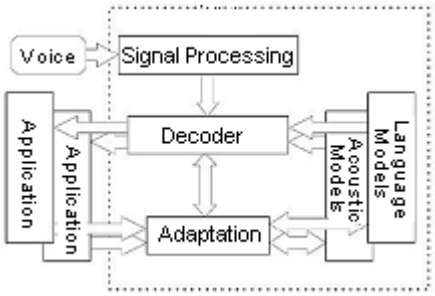

Figure 1. Basic Architecture of a Speech Recognition System

ACOUSTIC MODEL

The theory of acoustic phonetics postulates that there exists a distinctive, finite set of phonetic units in a language and that the phonetic units are broadly characterized by a set of properties that are manifested in the speech signal over time. Hence the first step in acoustic modeling is called segmentation and labeling because it involves segmenting the speech signal into discrete time intervals. This process occurs in the block labeled Adaptation of Figure 1.

Acoustic pressure waves are transformed into a description of the spectral characteristics of the speech signal using Digital Signal Processing (DSP) techniques. That description of the continuous speech waveform is then converted into a sequence of equally spaced discrete speech state vectors. It is then postulated that the duration (typically 10 ms) covered by a single speech state vector provides sufficient data to recover stable phoneme information. The speech state vectors shown in Figure 2 form an observation sequence. For convenience, we denote each of the vectors in the observation sequence by o_i where i = 1,2,…,n.

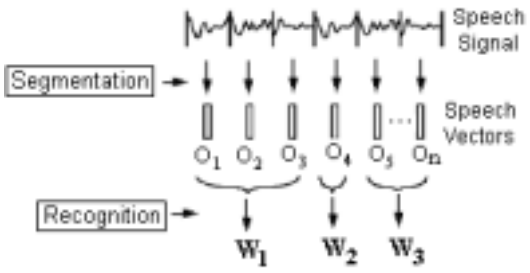

Figure 2. Acoustic Decoding

A probabilistic model of speech assumes that a specified word or word sequence, W, produces an acoustic observation sequence Y, with probability P(W|Y). The recognition problem can then be defined as determining the following:

$$\arg\max_{W} P(W|Y) \qquad (1)$$

This probability is not computable directly, but using Bayes' Rule, equation (1) can be rewritten as

$$P(W|Y) = \frac{P(Y|W)P(W)}{P(Y)} \qquad (2)$$

Since P(Y) is independent of W, the decoding rule becomes

$$\arg\max_{W} P(Y|W)P(W) \qquad (3)$$

The first term in equation (3), P(Y|W), is generally called the acoustic model, as it estimates the probability of a sequence of acoustic observations, conditioned on the word string.

A Markov model is a finite state machine that changes state at regularly spaced, discrete time units. Any system may be described as being in one of N distinct states. Transitioning from state i to j is probabilistic and is governed by a discrete probability a_{ij} that form a transition probability distribution $A = \{a_{ij}\}$. If one denotes the time instances associated with a state change as $t = 1,2,\ldots$ and the actual state at time t as q_t, then in a Markov chain we have as the probabilistic dependence between any state and its predecessor states given by

$$a_{ij} = P[\, q_t = j \mid q_{t-1} = i \,]; \quad 1 \le i,j \le N$$

where the right-hand side is considered to be independent of time.

In Hidden Markov Model (HMM) based speech recognition, it is assumed that a sequence of observed speech vectors corresponding to each word is generated by a Markov chain as shown in Figure 3. It is an example of a simple model with 5 states (labeled 1 to 5) where the entry and exit states are non-emitting. By traversing the state machine in this example, we produce an observation sequence. In practice, however, only the observation sequence is known and the underlying state sequence is hidden. Note that if all probabilities, a_{ij}, are nonzero, then it is possible to transition from any state directly to any other state. However, due to the nature of speech (i.e. a speech signal has properties such that it changes over time in a successive manner) an HMM as applied to speech recognition imposes certain constraints that govern the transition between states. Any prohibited transitions have a state transition probability $a_{ij} = 0$. Also, if one denotes that there may be M distinct observation symbols (e.g. phonemes) per state given by V_k, with $1 \le k \le M$, then the symbol probability distribution in state j is $B = \{b_j(o_k)\}$, in which

$$b_j(o_k) = P[\, o_t = V_k \mid q_t = j \,]$$

where o_t is a speech vector or observation at time, t, in the sequence $O = \{o_1, o_2, \ldots o_n\}$ having a total of n observations.

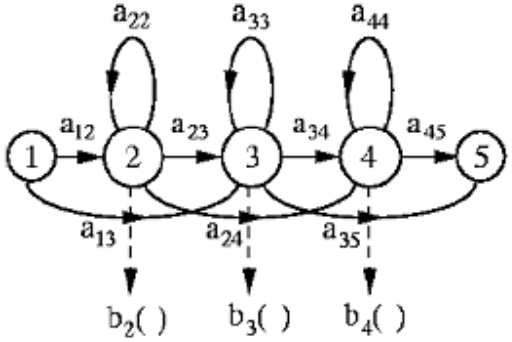

Figure 3. 5-State Left-to-Right Markov Chain

In computing P(Y|W), we need to use a statistical model for subword speech units. For our purposes, subword speech units consist of phonemes. The STG used phonemes with aggregation to form triphone units. Each vector o_i is a p-dimensional vector. Ideally, we have for each phoneme its single corresponding observation vector o_i. Practically speaking, more than one vector may correspond to one phoneme. In order to handle this situation a method called vector quantization (VQ) is used. These quantized vectors form a codebook, which is then used to find the most probable HMM and consequently the most probable corresponding word.

LANGUAGE MODEL

The second term in Equation (3), P(W), is generally called the language model, as it describes the probability associated with a postulated sequence of words. Such language models can incorporate both semantic and syntactic constraints of the language and the recognition task. If only syntactic constraints are used, the language model is called a grammar and may be viewed as a formal parser and syntax analyzer, N-gram word model (N = 2, 3, …), or a word pair grammar of some type. Generally such

language models are represented in a finite state network so as to be integrated into the acoustic model.

Unlike small vocabulary speech recognition systems that don't rely heavily on a language model to accomplish their selected tasks, a large vocabulary speech recognition system is dependent on linguistic knowledge, which can be presented in the form of a statistical language model. In a large vocabulary recognition system, a statistical language model provides an estimate of the probability of a word sequence W for the given recognition task. If we assume that W is a specified sequence of words, i.e.,

$$W = w_1 w_2 \dots w_m, \tag{4}$$

then P(W) can be computed as:

$$\begin{aligned} P(W) &= P(w_1 w_2 \dots w_m) \\ &= P(w_1) P(w_2|w_1) \\ &\times P(w_3|w_1 w_2) \dots P(w_m|w_1 w_2 \dots w_{m-1}) \end{aligned} \tag{5}$$

Since it is nearly impossible to estimate word probabilities, $P(w_m | w_1 w_2 \dots w_{m-1})$ for all word and word sequences possible in a language, we use an N-gram word model to approximate this term as:

$$P(w_j|w_1 w_2 \dots w_{j-1}) \approx P(w_j|w_{j-N+1} \dots w_{j-1}), \tag{6}$$

In other words, it is based only on the preceding N-1 words. It is computationally intensive to estimate N-gram probabilities when N is large. Hence, for practical purposes, we use N = 2 (bi-gram) or at most N = 3 (tri-gram). In practice, the binary indicator function that follows is used to specify which word pairs are valid in a bi-gram model.

$$P(w_j|w_k) = \begin{cases} 1 & \text{if } w_k w_j \text{ is valid} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Language models implemented as finite state networks can be integrated with an acoustic recognition decoding algorithm to provide efficient recognition.

Statistical Language Model and Its Perplexity

The language model, or probability of word sequences P(W), is essential for accuracy in large vocabulary speech recognition systems. Depending on the size of the vocabulary, it is impractical or impossible to explicitly define every possible sequence in a model, hence P(W) has to be estimated from a textual training corpus that is representative of the targeted domain of a language. In practice, the word sequence probability P(W) is approximated by an N-gram model as follows:

$$P_N(W) = \prod_{i=1}^{m} P(w_i|w_{i-1}, w_{i-2}, \dots, w_{i-N+1}), \tag{8}$$

where the conditional probabilities, $P(w_i \mid w_{i-1}, w_{i-2}, \dots w_{i-N+1})$ can be estimated using the simple relative frequency approach.

$$P(w_i|w_{i-1}, w_{i-2}, \dots w_{i-N+1}) = \frac{F(w_i, w_{i-1}, \dots, w_{i-N+1})}{F(w_{i-1}, \dots, w_{i-N+1})} \tag{9}$$

in which F is the number of occurrences of the string in the given training corpus. This implies that a relatively large corpus is needed to get a reliable estimate.

When crafting a language model, one should be interested in measuring how well the model will perform in speech recognition tasks. One approach to measuring performance is based on the concept of source of information in information theory. Suppose some source outputs sequences of words $(w_1, w_2, \dots, w_m)$ from a given vocabulary. Then the entropy of the source is defined as

$$\begin{aligned} H &= -\lim_{m\to\infty} \frac{1}{m} \sum_{w_1, w_2, \dots, w_m} (P(w_1, w_2, \dots, w_m) \times \\ &\quad \log_2 P(w_1, w_2, \dots, w_m)) \end{aligned} \tag{10}$$

This summation is over all possible sequences of words. If the source has statistical properties that can be completely characterized in a sufficiently long sequence that the source puts out, then the entropy can be computed as:

$$H = -\lim_{m\to\infty} \frac{1}{m} \log_2 P(w_1, w_2, \dots, w_m) \tag{11}$$

Since a typical sequence that approaches infinity is unattainable, we estimate entropy for a sufficiently large value of *m* as:

$$\hat{H} = \frac{1}{m} \log_2 P(w_1, w_2, \dots, w_m) \tag{12}$$

This estimate is feasible to evaluate and use as a metric of performance of a language model. One

interpretation of H from a speech recognition perspective is the degree of difficulty that the recognizer encounters in determining a word from the language model. However, it is more convenient to use perplexity of a source of information, which has a one-to-one relationship to entropy and gives a larger scaling difference between sources of information. Perplexity is computed as:

$$PP = 2^{\hat{H}} \tag{13}$$

or

$$PP = \hat{P}(w_1, w_2, \ldots, w_m)^{-\frac{1}{m}} \tag{14}$$

where

$$\hat{P}(w_1, w_2, \ldots, w_m)$$

is the probability estimate assigned to the word sequence $(w_1,w_2,\ldots,w_m)$ by a language model. Perplexity can be viewed as the average number of possible words following any string of N-1 words in a large corpus. It is often referred to as the average word branching factor of the language model.

AIR TRAFFIC CONTROL EXPERIMENTAL DATA ANALYSIS

In 2000, the Speech Technology Group (STG) at NAVAIR Orlando embarked on an endeavor to provide speaker independent, continuous speech recognition capabilities for the Navy Air Traffic Control (ATC) trainers located at NAS Pensacola. The STG was recruited for this task in order to upgrade an existing trainer that was equipped with hardware-based speech recognition capabilities and required each student to voice enroll prior to participation in a training exercise. The recognition performance was often less than optimal and required re-enrolling the student to try to correct the situation. With this in mind, the STG's objective was to build a software-based solution that would perform recognition for any student and could deliver at least 95% word accuracy under classroom conditions. The STG selected the HTK (Hidden Markov Model Tool Kit) developed by Cambridge Research Laboratories in order to build the speech models required by the Navy's ATC facility. The HTK toolkit employs a generic pattern recognition approach that can be applied to a variety of pattern recognition problems. Hidden Markov Model (HMM) based speech recognition is a statistical method of characterizing the spectral properties of the frames of a pattern (i.e. a spoken utterance). The HTK provides the ability to build models that perform continuous, speaker-independent speech recognition within a constrained domain. Further, it offers the ability to build custom acoustic models and language models versus exclusively building language models and utilizing generic acoustic models. The STG adopted this approach in order to more accurately model the acoustic patterns idiosyncratic of air traffic controllers. The vocabularies for ATC applications were large yet constrained and highly structured. Those vocabularies were divided into categories according to more specific types of air traffic control operations. The training facility at NAS Pensacola required the STG to build models that covered vocabularies for the Tower Operator Trainer System (TOTS), Carrier Air Traffic Control Center (CATCC) and the Amphibious Air Traffic Control Center (AATCC). Each of these vocabularies can be further categorized according to individual training positions for speech recognition purposes. For example, the CATCC trainer system has positions associated with departure, final, approach and marshal phases of aircraft handling as well as an Instructor Operator Station (IOS). This paper focuses on the CATCC Final position for analysis. The perplexity of the position yielded a value of 1.7397 when traversing the word network using 1,000,000 utterances. The maximum nodes required to traverse the network was 22. A traverse of the same network using only 1,000 utterances yielded a perplexity of 1.7446. This indicates that the perplexity is converging to a stable number (see Figure 4).

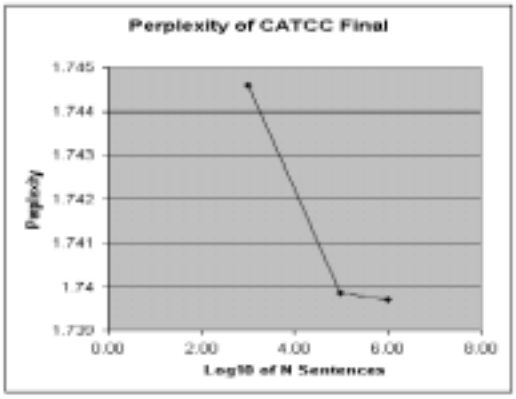

Figure 4. Analysis of perplexity

The construction of the network allowed an optional silence path between each word. A valid utterance consisted of a call sign (e.g. TOP HAT ONE SIX THREE) followed by a valid ATC phrase for that position. In addition, the network provided for a loop back to the beginning of a call sign or phrase from any word using the keyword CORRECTION at the speakers' discretion if a mistake was made. For comparison, a more complex position, CATCC IOS, yielded a perplexity of 1.8030 when traversing the word network with 1,000,000 phrases. The number of all possible utterances however far exceeds that of CATCC Final. In order to build the CATCC Final model, 10 individuals of the same gender each recorded approximately 4200 phrases at the training site. The CATCC IOS model was built with approximately four times the utterances of CATCC Final. The training utterances chosen ensured that every possible word pair was exercised at least once and every possible word was exercised at least twice. The word dictionary for final contained 216 words compared to 364 words for CATCC IOS, where the larger word count is reflected in the larger perplexity. The tools used to build the speech model provided for both a monophone and a triphone based model. The later was generated using an equivalent set of triphone transcriptions from the monophones transcriptions. The results of analysis of male gender models built for CATCC Final are illustrated in Figure 5 as follows:

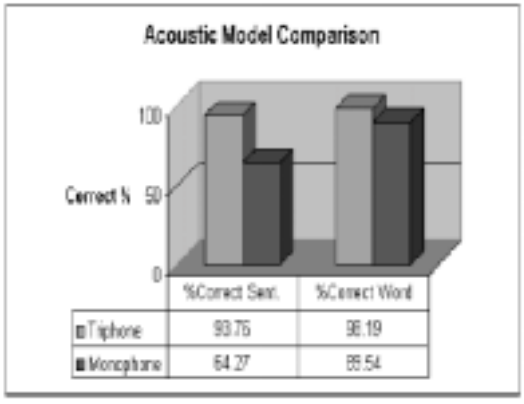

Figure 5. Acoustic models comparison

For comparison, the language model for the CATCC Final position was also built using Nuance's speech modeling toolkit. Nuance offers the ability to customize language models that can be used with their generic acoustic models. There are two approaches to designing language models with Nuance. In the first approach, the Grammar Specification Language (GSL) can be used to build models with highly constrained vocabularies. In a GSL model, all legal phrases must be explicitly defined in the model. In the second approach, a Statistical Language Model (SLM) is "trained" from a set of examples that models the users' speech rather than explicitly defining every legal phrase. To train an SLM grammar, a domain specific set of example phrases is passed to a Nuance utility that estimates the model probabilities. The CATCC Final position vocabulary was modeled using both approaches in Nuance. An analysis of the perplexity of the SLM using Nuance's product called "process-slm" yielded a value of 4.2918 when traversing the model using 1,000 phrases. Both the GSL and SLM were tested using 120 utterances from two speakers (one male and one female) for a total of 240 phrases collected under laboratory conditions. The results of those tests are illustrated in Figure 6.

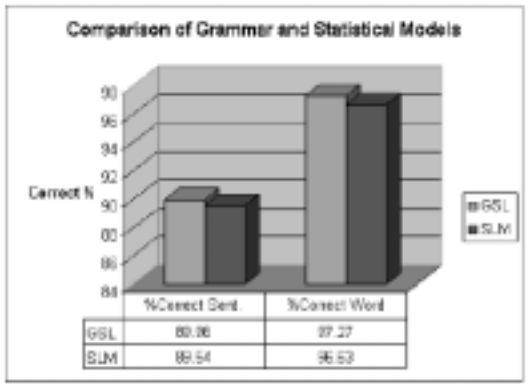

Figure 6. Language models comparison

CONCLUSION

Extensive tests of the custom acoustic and language models built with HTK show those models will perform with a rate of accuracy required for ATC applications. However, the process for building custom acoustic models is time consuming and requires undergoing the lengthy process of collecting and editing vast amounts of training data at the training site. Also, any minor change in vocabulary could require undergoing the costly process of collecting additional training data and potentially rebuilding the acoustic model. However, the process of building language models and utilizing generic acoustic models is far more rapid and does not require a data collection

process at the training site. The results from the preliminary test using the Nuance toolkit suggest that the language models could perform with a rate of accuracy sufficient for ATC applications. Consequently, an approach that forgoes building custom acoustic models for applications such as ATC merits further investigation.

REFERENCES

[1] F.K. Lee, C.H. Soong and B.H. Juang. A segment model based approach to speech recognition. In Proceedings of ICASSP 88, New York, NY, 1987.

[2] K.F. Lee. Automatic Speech Recognition – The Development of the SPHINX System. Kluwer Academic Publishers, Boston, 1989.

[3] L.R. Pieraccini R. Lee, C.H. Rabiner and J.G. Wilpon. Acoustic modeling for large vocabulary speech recognition. Computer Speech and Language, 1990.

[4] Lawrence Rabiner and Biing-Hwang. Juang. Fundamentals of Speech Recognition. PTR Prentice Hall, 1993.

[5] A.E. Rabiner L.R. Wilpon J.G. Rosenberg and D. Kahn. Demisyllable based isolated word recognition. IEEE Transaction on Acoustic, Speech, and Signal Processing, 1983.

[6] T. Svendsen and F.K. Soong. On the automatic segmentation of speech signals. In Proceedings of ICASSP 87, Dallas, TX, 1987.

[7] Hsiao-Wuen Hon Xuedong Huang Alex Acero and Raj Reddy. Spoken Language Processing: A Guide to Theory, Algorithm and System Development. Prentice Hall PTR, 2001.

[8] Steve Young and Julian Odell. HTK Book. Cambridge University, 1997.