DEVELOPING SPEECH RECOGNITION MODELS FOR USE IN TRAINING DEVICES

Speech Recognition Technology
R. Bradley Cope
Stephen G. Boemler
David Kotick
Naval Air Warfare Center
Orlando FL

I. INTRODUCTION

The Navy's 15G30 series of Air Traffic Control (ATC) trainers currently employ custom hardware speaker dependent speech recognition systems to replace human role players. Speaker dependent systems pay a penalty of about thirty minutes per one hundred words to enroll the trainee onto the system to develop a custom set of templates that work with the phraseology for an application. The 15G30 series of trainers use a vocabulary greater than five hundred words, and therefore, require approximately two and one half hours of acoustic enrollment. Further, the speaker dependent systems are typically ten percent more accurate than the speaker independent systems. It is hypothesized that the ten percent gap in accuracy can be narrowed by developing custom speaker independent finite state machine models using acoustic signals recorded in the environment in which the system will be used. Similarly, the acoustic data should be made up of unscripted utterances (which provides a byproduct of modeling the mannerism in speaking those utterances). Additionally, speaker independent solutions do not require training and can be implemented into an all software configuration using off-the-shelf hardware. The technology used in this effort is the Hidden Markov Model (HMM) statistical state machine. Intuitively, employing domain specific acoustic signals for the development of continuous speaker independent speech recognition should improve recognition accuracy because the variability in potential speech patterns will be bounded to the context of the phraseology for which the recognition system will be used. Each self-looping node in the model shown in Figure 1 is a sub-phonetic state. Word models are formed through concatenation of these sub-phonetic states. These models are created by recursively re-estimating the Gaussian distribution with observations for each set of possible feature vectors that occur in that state. States may be iterated for some observations, as shown above, for states two and four. The re-estimation merely adjusts the height and width of the distribution. The figure shows Gaussian distributions for a sequence of six observations. The re-estimation (called training) is accomplished by applying the Baum-Welch algorithm.

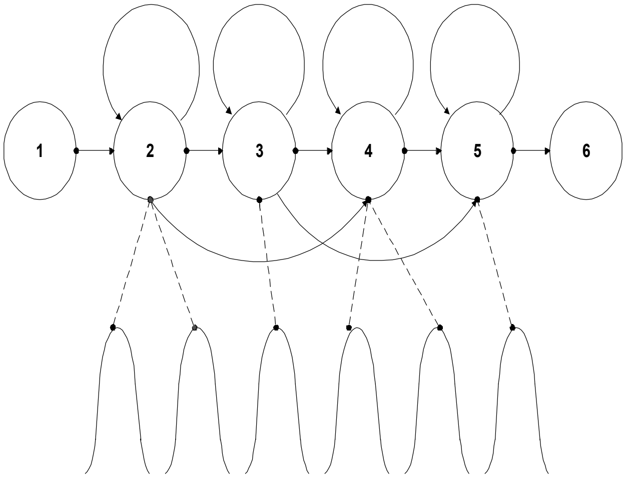

FIGURE 1 Markov Model

The variability of human speech is inherent to the Markov model after the model has been exposed to a representative set of subjects, each producing a set of utterances that will occur in the desired phraseology. Ideally, each possible utterance will have been spoken seven to ten times for each subject. A phonetic recognition system requires seven to ten occurrences of each phoneme in the context for which it will be used. Phonetic recognition, therefore, requires the number of phonemes (which is around forty for the English language), cubed because the phoneme can occur with many different surrounding phonemes thus changing the effects of coarticulation.

This paper describes a research effort leading to an improved speech recognition system for ATC training. We are attempting to make better finite state machine models using audio data recorded in the trainer where the recognition system is used. This data is intuitively better than commercial -off-the-shelf (COTS) model data, because COTS is designed for very large (60000 word plus) vocabularies where the models were typically created by people reading the Wall Street Journal (WSJ). Reading from a script is generally quite different than speaking naturally because of rhythm and emphasis. Further, the organization and frequency of occurrence for a set of words is dependent upon the application where the speech recognition system is to be used. The Wall Street Journal may have the words “final”, “radar”, and “contact” in the script, but a sequential occurrence of these words would be very unlikely. Additionally, there may be a certain rhythm that is used when speaking phrases, that is unique for a particular application, which would not be apparent in the non-domain specific scripted models. The phraseology used by Navy ATC controllers in training has been captured onto digital audio tape which can be used to develop domain specific HMM’s. The recorded data reflects the naturally occurring rhythm and context of the phraseology for the application as well as the acoustic environment for which the speech recognition system will be used.

II. BACKGROUND

Speech recognition is signal processing and modeling of spoken input for the communication of information. There are a few flavors of speech recognition depending on the technology used to identify a given utterance. The technology used in this exploratory research effort is speaker independent, continuous, speech recognition using Markov Modeling which is a probabilistic pattern matching approach that models a time-sequence of speech patterns representing the temporal structure of a phoneme or word depending on the type of recognizer used, phoneme based or word based. Continuous speech recognition is challenging because of the effects of coarticulation between words when words are strung together. The coarticulation of words is unique for domain specific phraseologies. The state of the technology is such that applications of continuous speech recognition will work reliably for small vocabularies. Modeling of the coarticulation effects of these words can be bounded because the vocabulary is small enough to allow for the building of models re-estimated with sufficient repetition of possible utterances. Further, re-estimation of these models with data from many different speakers, provides inherent flexibility of the models to allow for the inherent variability of spoken language[1]. The process of re-estimation effectively adjusts the Gaussian distribution of the possible feature vectors that can occur during the states in the Markov model[2]. Models adjusted for the phraseology and the mannerism in which a certain group of people communicate is an optimization of the model.

Speech recognition begins by sampling an analog microphone input with an analog-to-digital converter (A/D). The sampling rate is at least twice the highest signal frequency, commonly known as the Nyquist frequency, which prevents aliasing of the sampled signal. The digital audio is then transformed from the time domain to the frequency domain by way of a Fast Fourier Transform (FFT). These transforms are performed periodically on the input using a Hamming window. The bandwidths of the frequency components are based on the biologically inspired mel scale which has more resolution at the lower frequencies. Subsequently, the spectrum is run through a series of cosine functions to characterize the cepstral energy thus obtaining the mel cepstral coefficients (MFC’s). We use ten to twenty millisecond windows because of the mechanical operation of the articulatory components, especially the glottis, and it is assumed that this time period is short enough for the signal to be stationary. Each of the feature vectors in this experiment represent a ten millisecond sample. Hidden Markov Models (HMM) are developed by re-estimation of each possible state and establishing a distribution of the MFC classifications that could occur for each ten millisecond window. These models use a feed forward state transition topology to model the transitions between each sub-phonetic window. The Viterbi[3], or Baum-Welch re-estimation algorithms, then compute the statistical likelihood of the model producing a given spoken input or sequence of sub-phonetic observations.

III. DEVELOPMENT OF FINITE STATE MACHINE MODELS

Development and integration of speech recognition Markov models is simplified using a toolkit such as the Hidden Markov Model Toolkit (HTK) by Entropic. This toolkit operates on a Sun computer and requires one to know how to write shell scripts for automating the model development. Shell scripts are used to semi-automate most of the process from taking recorded audio and to convert it to the correct Sun file format by way of a data link with a SCSI bus connection. Further, scripts are used to manipulate the data for developing the models. Once created and trained, the models are used for recognition in the Viterbi recognition algorithm.

Finite state machine HMM's are partitioned phonetically or lexically. When the partitioning is phonetic, words are constructed by concatenating the phonetic based models together. Each ten millisecond state of the phonetic model has a probability distribution for the feature vectors that can occur for that moment in time. Initially, the probability distribution is established by aligning the acoustic signal with a prescribed phonetic topology for the expected word. Subsequently, the probability distribution is set by re-estimating a large set of feature vectors specific to the phraseology from a variety of human subjects. The prescribed phonetic topology is defined in a phonetic dictionary. This dictionary can include many variations of a given word which means there will be a unique set of phonemes for each possible variation. Air traffic control phraseology has unique concatenation of words and therefore, unique effects of coarticulation.

The original phonetic topology for a word is often changed and/or lost when spoken with surrounding words. It is this change and/or loss of phonetic information that establishes a unique and context specific vocabulary for the air traffic control application. For example, the coarticulation effects when concatenating the words "plugged" "and" "receiving" results in a loss of the beginning and the end of the word "and" leaving just the "n" as well as a change in the suffix "ed" which sounds more like a "t". Further, there is a change in the temporal topology of the model because it takes less time to pronounce the concatenated set of words than it does to pronounce each word individually. Figures 2 and 3 show, in the time domain and in the frequency domain, the words "plugged" "and" "receiving" spoken individually. Figures 4 and 5 show the effects of coarticulation with the concatenation of the words "plugged and receiving" in the time domain and in the frequency domain. It is clear that the concatenation of these words results in the loss of phonetic information.

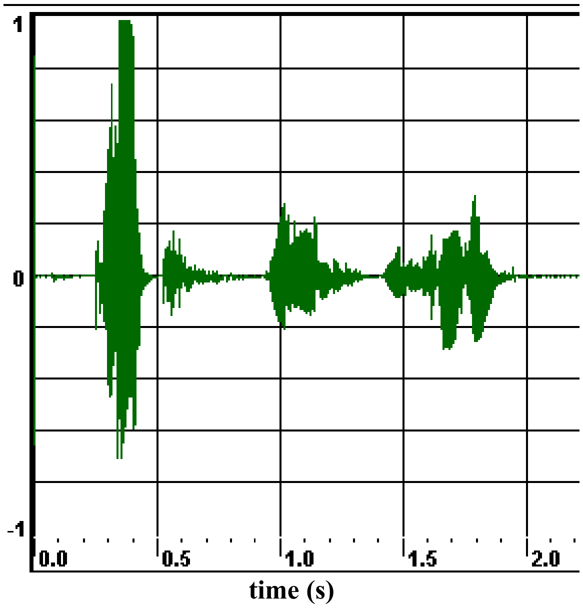

FIGURE 2

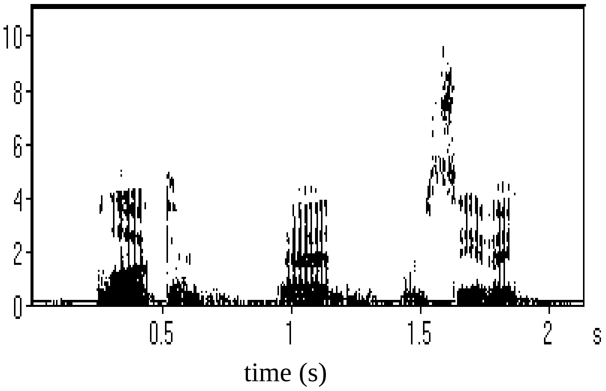

FIGURE 3

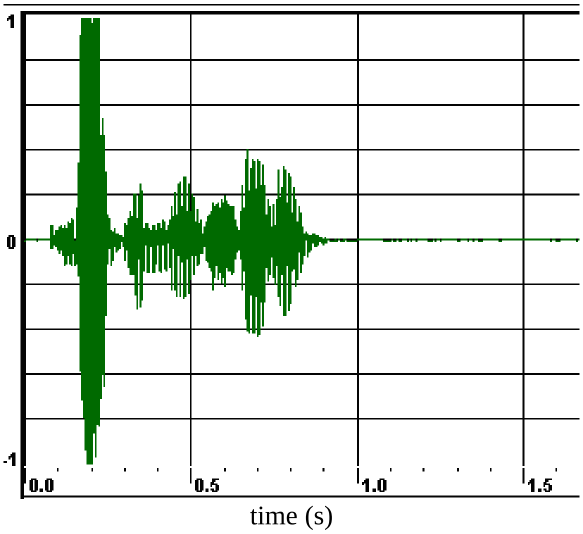

FIGURE 4

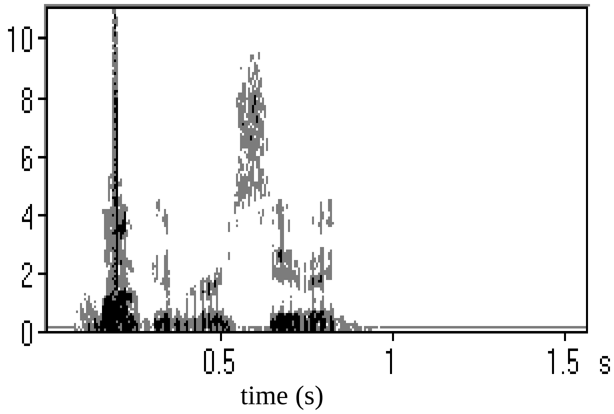

FIGURE 5

A gain in accuracy is realized when the phonetic dictionary reflects the effects of coarticulation thus creating a optimal set of word models for a given phraseology. Simply identifying and labeling in the dictionary, phonetic topology for a given unique utterance, establishes the alignment of the training data with the HMM's. The language model identifies the lexicon and syntax of the phraseology. Lexically and syntactically bounding the phraseology limits the combinations of possible phonetic concatenations. Further, the language model provides a lexical transition probability. This lexical probability combined with the HMM state probability results in the overall probability for a given utterance[4]. Use of models that are not optimized for the target phraseology results in a solution that can never perform at the limit of the technology. It is hypothesized that inclusion of the unique coarticulation effects of concatenated words leads to the best technical solution because the dictionary reflects unique pronunciations and the models are trained with data representing the actual usage of the phraseology.

Ideally, the models would be trained using acoustic signals that were recorded in the environment where the recognizer will be used. Similarly, it is best to train the models using the same microphone as that used in the target environment. Both of these approaches reduce the variability in the recognition of live audio input. Figure 6 shows the addition of 12 dB of noise to the phrase "plugged and receiving". White noise was added to test the response of the speech recognition system.

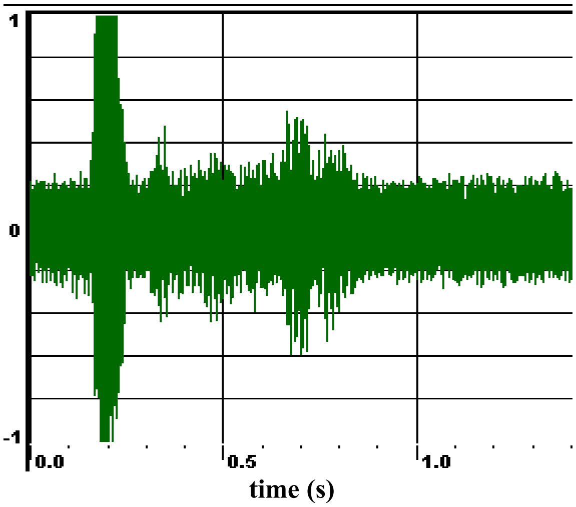

FIGURE 6

IV. RESULTS

The ATC HMM's were trained with acoustic signals comprised of Navy ATC trainees using the trainer during training exercises. These signals capture the emphasis, speed, and idiomatic characteristics particular to Naval ATC students. The ATC HMM's were trained with acoustic signals recorded in the target environment with the same input device that will be used with the new models thus reducing some of the environmental and equipment variability and therefore an optimized model. Testing the models presents a challenge because of the enormous possible combinations of valid utterances within the boundaries of the language model. A set of utterances representing most of the frequently used utterances was established as test data and run through the recognition system. Simply speaking a canned set of utterances to the recognition system will not provide repeatability or the rhythm, speed, emphasis, etc. that is inherent with data extracted from the acoustic signals recorded in the target environment.

Preliminary results for a small vocabulary implementation indicate that the optimized models perform better than the models developed from scripted and non-domain specific speech. Models for the digits zero through nine were trained with data from three adult male subjects. The models were then tested with a canned data set representing the target environment with the coarticulation specific to the target phraseology. The recorded data was verified with the transcriptions. Subsequently, the recorded data was run through both the domain and non-domain specific models. The domain specific models are labeled NAWC and the non-domain specific labels are labeled WSJ in the following tables. Further, two flavors of acoustic signals were tested, one with just the ambient background noise and the other mixed with 12 dB of white noise. Tables 1 and 2 show the results for the models tested with the ambient noise and tables 3 and 4 show the results for the models tested with an additional 12 dB of noise added to the recorded speech. Correctness does not include the insertions whereas accuracy does include the insertions. Tables 1 and 3 show the correctness and tables 2 and 4 show the accuracy. The same data set was tested against each model.

Correctness w/Ambient Noise

Model	W	S	D	Correct
NAWC	1865	5	3	99.57%
WSJ	1865	9	1	99.46%

TABLE 1

Accuracy w/Ambient Noise

Model	W	S	D	I	Accuracy
NAWC	1865	5	3	4	99.36%
WSJ	1865	9	1	2	99.36%

TABLE 2

Correctness w/12 dB of White Noise

Model	W	S	D	Correct
NAWC	1865	519	86	67.56%
WSJ	1865	241	302	70.56%

TABLE 3

Accuracy w/12 dB of White Noise

Model	W	S	D	I	Accuracy
NAWC	1865	519	86	55	64.61%
WSJ	1865	241	302	242	57.59%

TABLE 4

W-Words S-Substitutions D-Deletions I-Insertions

The results were scored using the National Institute for Science and Technology figure-of-merit methodology[5]. Basically there were a total of 1865 words in the data set. Substitutions occur when a word other than what was actually spoken is recognized. A deletion is an omission of a correctly spoken word. An insertion is anytime a word was added to a recognized utterance. The percentage of the correct words is the total number of words minus the substitutions and deletions divided by the total number of words. Accuracy is the total number of words minus the substitutions, deletions, and insertions all divided by the total number of words. Insertions or deletions of silence were omitted from the results because insertion and or deletion of silence does not affect the accuracy of the system. Substitutions of silence were not omitted from the results because a substitution of silence does affect the accuracy of the system.

The metric for correct words does not include word insertions which are mostly due to noise. Correctness is important when looking for a specific key word within a string. Accuracy includes the adverse effect of word insertions or forced recognition due to noise. Noise can be mitigated through signal processing and microphone improvements using arrays, but insertions due to noise must be considered, therefore the metric of accuracy reflects the performance of the recognition system.

V. CONCLUSIONS

This paper describes a pilot test of the process used to develop Hidden Markov models for speech recognition. The models were based on a subset of the air traffic control trainer vocabulary. The pilot test was necessary to establish the process for testing larger vocabularies. Further, this test provides a data point of the NAWC models. Development of these models is an iterative process. Exposure to new data enhances the variability of the models and the models will need to be tested periodically during development using the process established in this pilot test. The results, including the noise, show the NAWC domain specific models are an improvement over off-the-shelf models built with WSJ data, although more data is necessary to verify this conclusion. Any small improvement in speech recognition accuracy is significant.

VI. FUTURE WORK

This is a three year research effort and we are in the fourth quarter of the first year. The vocabulary for this project is considerably larger than the vocabulary used for the pilot test described in this paper. The Markov models for the pilot test were word based models. The models for the complete system may be word based models, or they may be phoneme based models. Phonetic models provide a higher context resolution because they are smaller in size. Both types of models will be constructed and evaluated for accuracy and correctness. Further, the models will be tested with representative noise from the target environment added to the test data.

VII. REFERENCES

1) Owens, F.J., *Signal Processing of Speech* , McGraw Hill, 1993

2) Young, Steve, *The Hidden Markov Model Toolkit ,* Cambridge University, 1997

3) Viterbi, A.J. “Error Bounds for Convolution Codes and Asymptotically Optimum Decoding Algorithm” IEEE Transactions on Information Theory, Vol IT-13, April 1967, pp260-269.

4) Ravishankar, Mosur K. *Efficient Algorithms for Speech Recognition* , Carnegie Mellon University, May 15, 1996

5) ”*Figure of Merit for the Evaluation of Speech Recognition* , National Institute for Science and Technology